NTNU
Norwegian University of
Science and Technology

Methods for Extreme Value Statistics Based on Measured Time Series

Even Haug

Master of Science in Physics and Mathematics
Submission date: June 2008
Supervisor: Arvid Næss, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

The master project focuses on the development of a general method for extreme value estimation based on sampled time series. The method is designed to account for statistical dependence between peak values in a rational way. This avoids the problem of declustering of data to ensure independence, which is a common problem for e.g. the peaks over threshold method. The goal is to establish an accurate method for prediction of e.g. extreme wind speed based on observation data. The method will be tested on real data.

Assignment given: 21. January 2008
Supervisor: Arvid Næss, MATH

Preface

This work is a continuation of my project work, *Estimering av ekstremverdier ut fra målte dataserier ved topp over terskel- og AER-metoden* ("Estimation of Extreme Values from Measured Data Series by the 'Top over Threshold' and the AER Methods"), which was written during the fall semester of 2007 at NTNU, in Norwegian.

I want to thank Professor Arvid Næss, who has been my supervisor during the last two semesters at NTNU, and Øyvind Breivik at the Norwegian Meteorological Institute, who has provided some of the data used in this work.

Even Haug
June 12, 2008
Gløshaugen, Trondheim

Abstract

The thesis describes the Average Exceedance Rate (AER) method, which is a method for predicting return levels from sampled time series. The AER method is an alternative to the Peaks over threshold (POT) method, which is based on the assumption that data exceeding a certain threshold will behave asymptotically. The AER methods avoids this assumption by using sub-asymptotic data instead. Also, instead of using declustering to obtain independent data, correlation among the data is dealt with by assuming a Markov-like property.

A practical procedure for using the AER method is proposed and tested on two sets of real data. These are a set of wind speed data from Norway and a set of wave height data from the Norwegian continental shelf. From the results, the method appears to give satisfactory results for the wind speed data, but for the wave height data its use appears to be invalid. However, the method itself seems to be robust, and to have certain advantages when compared to the POT method.

Contents

Preface **i**

Abstract **iii**

1 Introduction **1**

2 Theory **2**
2.1 Return Levels 2
2.2 The POT Method 2
2.3 The Average Exceedance Rate (AER) Method 4
2.4 Estimation of $\bar{\epsilon}_k(\eta)$ 7
2.5 Estimation of the Constants 9
2.6 Step 1 11
2.7 The Numerics 15
2.8 Step 2 16
2.9 Weights 18
2.10 Confidence Intervals 18

3 The Data **20**
3.1 The Wind Speed Data 20
3.2 The Ocean Wave Data 24

4 Results **28**
4.1 The Ørlandet Wind Speed Data 28
4.2 The Alta Wind Speed Data 41
4.3 The Ocean Wave Data 46
4.4 Discussion 54

5 Conclusion **60**

A R Code **63**

1 Introduction

Extreme value statistics is the branch of statistics that deals with unusual events, such as the very smallest or the very greatest levels of a process. The discipline may be used to estimate the risk of an unusual event occurring or the maximum value of a physical quantity during a long time span. It is therefore of great use to engineers, who want to estimate the magnitude of the forces that may be expected to affect a structure. Especially, one will often be interested in the long-term return levels of the extreme values. The return levels are the levels one expects to be exceeded by the process during a certain time interval.

When predicting the future, we must rely on information from the past. But since unusual events are scarce and belong to the very tail of the distribution of the phenomenon under study, we may not have observations on the levels we are interested in. In extreme value statistics, this problem is overcome by fitting the tail of the distribution and extrapolating from the known levels to the unknown. In particular, it is often assumed that the extreme values among the observed events belong to the asymptotic part of the tail, and that an extreme value distribution can be fitted to those data. This assumption is the foundation of both the *Generalized Extreme Value* (GEV) method and the *Peaks over Thresholds* (POT) method, both of which are widely used in practice. There is an extensive literature on the subject, for example [2].

The GEV and POT methods rely on the assumption of the data used in the analysis stemming from the asymptotic part of the distribution tail. In practice, however, it is impossible to ascertain whether this is true or not, and so we cannot know if the assumption is valid and the methods are useable. Using them requires, it has been noted, "a leap of faith" [2, p. vii]. Therefore, an alternative method has been developed, a method that relies also on data from the subasymptotic part of the distribution, and hopefully will prove to be more flexible and robust. This is the *Average Exceedance Rate* (AER) method.

The AER method is a fairly recent method. Except for [11], the literature on the subject is scarce, and there has not been done much work on the practical implementation of the method. In this work, after describing the theoretical background of the method, we will propose a practical implementation for extreme value estimation based on sampled time series. Especially, we must then account for the statistical dependence between the data. The method will be tested on four sets of real data. The two first consist of 7 years of daily wind speed observations from Ørlandet Airport and Alta Airport in Norway. The two last consist of almost 13 years of ocean wave height observations from the Draugen and Ekofisk oil fields off the coast of Norway. The analysis of the Ørlandet wind speed data will be thoroughly discussed in order to illustrate the practical use of the method, while the results from the other data sets will be presented and compared.

In section 2, the theoretical background of the method and a proposed practical implementation of it are presented. In section 3, the wind speed and ocean wave data are presented and discussed. Then, in section 4, a detailed practical example of using the AER method is given, before the results are presented and discussed. Finally, in

section 5, there are some concluding remarks. To illustrate the actual implementation of the method on a computer, some of the R code used in the analysis is added in appendix A.

2 Theory

We shall consider a time series of observations $X = \{X_1, X_2, \ldots, X_N\}$ derived from a stochastic process, where the distribution of the elements $X_j, j = 1, \ldots, N$ are considered unknown, and where the observations were made at discrete times $t_j, j = 1 \ldots, N$ over a time span of length T. The extreme value of the series, M_N, is the largest among the elements,

$$M_N = \max\{X_j; j = 1, \ldots, N\}. \tag{1}$$

The cumulative distribution of the extreme value, $\text{Prob}(M_N \leq \eta)$, will be referred to as $P(\eta)$.

2.1 Return Levels

Now, if we let the time span $(0, T)$ of the time series represent a period, we are interested in finding the return level x_m, the level which we expect the process to exceed every m periods. In other words, the return level is the level which for every period will be exceeded with a probability of $\frac{1}{m}$.

An element in the time series exceeding the level x_m is equivalent to the extreme value M_N of the series exceeding x_m. We can therefore relate the exceedance probability to the cumulative distribution $P(\eta)$ of the extreme value M_N, setting

$$\text{Prob}(M_N > \eta = x_m) = 1 - \text{Prob}(M_N \leq \eta = x_m) = \frac{1}{m}. \tag{2}$$

In practice, a convenient time span such as one year is often used as the period.

2.2 The POT Method

A widely used method to estimate the return levels is the POT method. In order that we may understand the differences between the AER method and the POT method, a brief exposition of the latter will here be given, before we proceed to the AER method. The exposition follows [2].

Since the extreme value M_N is the greatest among the elements of X, it will belong to the very tail of the distribution of those elements. Now, since we do not know the distribution, we cannot say anything about the tail. However, it can be shown that asymptotically, we can actually say something about the distribution of M_N. When $N \to \infty$, the distribution of M_N will converge against one of only three extreme value distributions. Those are the Gumbel distribution, the Fréchet distribution and the Weibull distribution. All three distributions can be written on the form of the generalized extreme value distribution,

$$G(\eta) = \exp\left\{-\left[1+\xi\Big(\frac{\eta-\mu}{\sigma}\Big)\right]^{-\frac{1}{\xi}}\right\}, \tag{3}$$

where μ is a location parameter, $\sigma > 0$ is a scale parameter, and ξ is a shape parameter. The case $\xi > 0$ is equivalent to the Fréchet distribution, the case $\xi > 0$ is equivalent to the Weibull distribution, and the case ξ is equivalent to $G(\eta), \xi \to 0$, which gives the Gumbel distribution,

$$G(\eta) = \exp\left\{-\exp\Big[-\Big(\frac{\eta-\mu}{\sigma}\Big)\Big]\right\}. \tag{4}$$

In short, we have

$$P(\eta) = \text{Prob}(M_N \leq \eta) \to G(\eta), N \to \infty. \tag{5}$$

If (5) is valid, it can also be shown that for a high threshold u, the residuals $Y = X - u$ follow the distribution

$$\text{Prob}\{Y = y | X > u\} = H(y) = 1 - \Big[1+\xi\Big(\frac{y}{\tilde{\sigma}}\Big)\Big]^{-\frac{1}{\xi}}, \tag{6}$$

where

$$\tilde{\sigma} = \sigma + \xi(u-\mu), \tag{7}$$

and σ, μ, and ξ are the parameters of the corresponding GEV distribution of M_N. This distribution is called the generalized Pareto (GP) distribution. So, if M_N follows a GEV distribution, the residuals $X - u$ follow a GP distribution with the same parameters as in the GEV distribution.

The foundation of the POT method is the assumption of using asymptotic data. That is, we assume that the approximation $P(M_N \leq \eta) \approx G(\eta)$ is valid for large N and for η values larger than some high threshold u. A GP distribution is then fitted to the residuals $\{X - u | X > u\}$, using the maximum likelihood method to find estimates of the constants μ, σ and ξ. Estimates of the return levels are calculated from (2) using the estimated extreme value distribution. The exceedances of the threshold are assumed to be independent, and if in practice they are not, one can extract independent exceedances by declustering.

The weakness of the POT method is the very assumption of using asymptotic data. In practice, it is impossible to determine if the approximation is valid, so it does really rest on a "leap of faith" [2, p. vii] [11]. We would assume that N as well as the number of exceedances of the threshold would have to be large for the assumption to be valid, but in practice, we will often have very few exceedances of the highest levels to work with [3]. Using declustering, many of the observations will be left out from the analysis. Hence, if there is great uncertainty among the data, the estimates will not be very reliable. Indeed, one or two outliers among the greatest values may distort the return level estimates greatly.

2.3 The Average Exceedance Rate (AER) Method

To overcome the problem of having to assume asymptotic data, the Average Exceedance Rate (AER) method has been developed. This method also takes into account data from the sub-asymptotic part of the distribution, so that more data may be used and the estimates are less dependent on the greatest values. It has a limitation, though, in that it is assumed that the extreme values really follow a Gumbel distribution.

In this and in the next section, the theoretical background of the AER method will be exposed. The exposition mainly follows [11]. The theoretical foundation of the method is also discussed in [12] and [13].

We want to find a way of estimating the return levels x_m from the time series X. Now, as we have seen, the return levels are related to the cumulative distribution $P(\eta)$ of the extreme values. Hence, we can estimate the return levels by way of estimating $P(\eta)$. To be able to estimate $P(\eta)$, we will have to make a few assumptions. First, we need to take into account the correlation among the data in the time series. To use the POT method, we had to assume independence among the exceedances of the threshold, or else use declustering to select only independent exceedances. This will possibly leave out a great number of data points which could have been useful in estimating the GP distribution. Using the AER method, however, it is not necessary to assume independence among the data. Instead, we assume that our time series has a Markov-like property. Given the time series $X = \{X_1, \ldots, X_N\}$, we have

$$\begin{aligned} P(\eta) &= \text{Prob}\{X_1 \leq \eta, \ldots, X_N \leq \eta\} \\ &= \text{Prob}\{X_N \leq \eta | X_1 \leq \eta, \ldots, x_{N-1} \leq \eta\} \text{Prob}\{X_1 \leq \eta, \ldots, X_{N-1} \leq \eta\} \\ &= \prod_{j=2}^{N} \text{Prob}\{X_j \leq \eta | X_1 \leq \eta, \ldots, X_{j-1} \leq \eta\} \cdot P(X_1 \leq \eta). \end{aligned} \tag{8}$$

Introducing the Markov-like property, we write

$$\text{Prob}\{X_j \leq \eta | X_1 \leq \eta, \ldots, X_{j-1} \leq \eta\} \approx \text{Prob}\{X_j \leq \eta | X_{j-k+1} \leq \eta, \ldots, X_{j-1} \leq \eta\} \tag{9}$$

for a suitable k. That is, we assume that the element X_j of the time series is only dependent on the $k-1$ last elements. If $k = 1$, the elements are independent. For convenience, we will set $k = 2$ in the following exposition, otherwise referring to [11]. Then, we will have

$$
\begin{aligned}
P(\eta) &\approx \prod_{j=2}^{N} \text{Prob}\{X_j \le \eta | X_{j-1} \le \eta\} \cdot P(X_1 \le \eta) \\
&= \prod_{j=3}^{N} \text{Prob}\{X_j \le \eta | X_{j-1} \le \eta\} \cdot \text{Prob}\{X_2 \le \eta | X_1 \le \eta\} \cdot P(X_1 \le \eta) \\
&= \prod_{j=3}^{N} \text{Prob}\{X_j \le \eta | X_{j-1} \le \eta\} \cdot \text{Prob}\{X_1 \le \eta, X_2 \le \eta\} \\
&= \frac{\prod_{j=2}^{N} \text{Prob}\{X_{j-1} \le \eta, X_j \le \eta\}}{\prod_{j=3}^{N} \text{Prob}\{X_{j-1} \le \eta\}} = \frac{\prod_{j=2}^{N} \text{Prob}\{X_{j-1} \le \eta, X_j \le \eta\}}{\prod_{j=2}^{N-1} \text{Prob}\{X_j \le \eta\}} \\
&= \frac{\prod_{j=2}^{N} p_{2j}(\eta)}{\prod_{j=2}^{N-1} p_{1j}(\eta)},
\end{aligned} \tag{10}
$$

where we have introduced the notation

$$
p_{kj}(\eta) = \text{Prob}\{X_{j-k+1} \le \eta, \ldots, X_j \le \eta\}, j \ge k. \tag{11}
$$

Next, we introduce the expression

$$
\alpha_{kj}(\eta) = 1 - \frac{p_{kj}(\eta)}{p_{k-1,j-1}(\eta)}, j \ge k \ge 2. \tag{12}
$$

which can be written

$$
\alpha_{kj}(\eta) = \text{Prob}\{X_j > \eta | X_{j-k+1} \le \eta, \ldots, X_{j-1} \le \eta\}, \tag{13}
$$

and which is an expression for the probability that the element X_j will exceed the threshold η, supposing the $k-1$ last elements in the time series did not exceed that threshold. Let us call such an exceedance a "conditional exceedance" of η. An alternative way of writing $P(\eta)$ in (10) will then be

$$
P(\eta) \approx \frac{\prod_{j=2}^{N} p_{2j}(\eta)}{\prod_{j=2}^{N-1} p_{1j}(\eta)} = \prod_{j=2}^{N} \frac{p_{2j}(\eta)}{p_{1,j-1}(\eta)} \cdot p_{1N}(\eta) = \prod_{j=2}^{N} \Big(1 - \alpha_{2j}(\eta)\Big) \cdot p_{1N}(\eta). \tag{14}
$$

For $k = 1$, we define

$$
\alpha_{1j} = \text{Prob}\{X_j > \eta\} = 1 - p_{1j}(\eta), \tag{15}
$$

which gives

$$
P(\eta) \approx P_1(\eta) = \exp\Big(-\sum_{j=1}^{N} \alpha_{1j}(\eta)\Big). \tag{16}
$$

If we approximate $1 - \alpha_{kj}(\eta) \approx \exp\{-\alpha_{kj}(\eta)\}$, for $k = 2$ we can write

$$P(\eta) \approx P_2(\eta) = \exp\Big(-\sum_{j=2}^{N} \alpha_{2j}(\eta) - \alpha_{1N}(\eta)\Big) \approx \exp\Big(-\sum_{j=2}^{N} \alpha_{2j}(\eta)\Big), \tag{17}$$

if N is large. Generally, we can write [11]

$$P(\eta) \approx P_k(\eta) \approx \exp\Big(-\sum_{j=k}^{N} \alpha_{kj}(\eta)\Big). \tag{18}$$

Let us investigate more closely what the sum $\sum_{j=2}^{N} \alpha_{kj}(\eta)$ expresses. For $k \geq 2$, $\alpha_{kj}(\eta)$ is the probability that X_j is a conditional exceedance. The sum $\sum_{j=1}^{N} \alpha_{kj}(\eta)$ therefore is the expected number of conditional exceedances of the level η during the time span $(0, T)$. Similarly, $\sum_{j=1}^{N} \alpha_{1j}(\eta)$ is the expected number of unconditional exceedances.

We now introduce the average exceedance rate (AER) $\bar{\epsilon}_k(\eta)$, which is expressed as

$$\bar{\epsilon}_k(\eta) = \frac{1}{N-k+1} \sum_{j=k}^{N} \alpha_{kj}(\eta), \approx \frac{1}{N} \sum_{j=k}^{N} \alpha_{kj}(\eta), k = 1, 2, \ldots \tag{19}$$

The expression in (18) can then be rewritten

$$P_k(\eta) \approx \exp\{-\bar{\epsilon}_k(\eta) N\}. \tag{20}$$

Thus, it is possible to estimate $P(\eta)$ by way of $\bar{\epsilon}_k(\eta)$. Of course, as long as the underlying distribution of the elements in the time series is unknown, we do not have an expression for $\bar{\epsilon}_k(\eta)$. But asymptotically, we know that $P(\eta)$ must converge to one of the Weibull, the Gumbel or the Fréchet distributions. Especially, if we assume that our underlying distribution converges to the Gumbel distribution, by comparing (4) and (20) we find that asymptotically, $\bar{\epsilon}_k(\eta) \propto \exp\{-a(\eta - b)\}$, where b corresponds to μ of the Gumbel distribution and a corresponds to $\frac{1}{\sigma}$.

Now, we assume that sub-asymptotically, the average exceedance rate can be approximated by

$$\bar{\epsilon}_k(\eta) \approx q(\eta) \exp\{-a(\eta - b)^c\}, \eta \geq \eta_1, \tag{21}$$

where the function $q(\eta)$ varies slowly compared to the exponential function, and a, b, c are the above-mentioned constants. It is assumed that this approximation is valid for η larger than some η_1. For $c = 1$, (21) is of course equivalent to the asymptotic Gumbel distribution.

We further assume that for large η, $q(\eta)$ will vary so slowly that it can be approximated by a constant q. Then (21) can be rewritten

$$\bar{\epsilon}_k(\eta) \approx q \exp\{-a(\eta - b)^c\}, \eta \geq \eta_1. \tag{22}$$

Now, combining (2) and (20), and inserting the approximation (22), we can find an expression for the m periods return values,

$$\begin{aligned} P_k(M_N > \eta = x_m) = 1 - P_k(\eta = x_m) &= 1 - \exp\{-\bar{\epsilon}_k(x_m)N\} = \frac{1}{m} \\ -\bar{\epsilon}_k(x_m)N &= \log\big(1 - \frac{1}{m}\big) \\ q\exp\{-a(x_m - b)^c\}N &= -\log\big(1 - \frac{1}{m}\big) \\ -a(x_m - b)^c &= \log\Big(-\frac{\log(1 - \frac{1}{m})}{qN}\Big) \\ x_m &= \Big[-\frac{1}{a}\log\Big(-\frac{\log(1 - \frac{1}{m})}{qN}\Big)\Big]^{\frac{1}{c}} + b. \end{aligned} \quad (23)$$

Thus, it turns out that we can estimate the return levels by finding suitable estimates of the constants q, b, a, and c in the approximation of $\bar{\epsilon}_k(\eta)$.

2.4 Estimation of $\bar{\epsilon}_k(\eta)$

To find estimates of the constants q, b, a, and c, we first have to find estimates $\hat{\epsilon}_k(\eta)$ of the average exceedance rate $\hat{\epsilon}_k(\eta)$. The average exceedance rate can be written as

$$\bar{\epsilon}_k(\eta) = \mathrm{E}[\beta_k(\eta)], \quad (24)$$

where

$$\beta_k(\eta) = \frac{1}{N - k + 1}\sum_{j=k}^{N} \mathbf{1}_{kj}(\eta), \quad (25)$$

and where $\mathbf{1}_{kj}$ is an indicator variable for the event $\{X_j > \eta | X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}$. For $k = 1$, estimating $\beta_1(\eta)$ will be particularly simple, since for each element X_j of the time series we do not condition on any previous elements. Our estimate will be

$$\hat{\beta}_1(\eta) = \frac{1}{N}\sum_{j=1}^{N} I(X_j > \eta), \quad (26)$$

where $\sum_{j=1}^{N} I(X_j > \eta)$ is the counted number of exceedances of the level η. However, using (25) to estimate $\beta_k(\eta)$ for $k \geq 2$ is impractical. Instead, we note that we can write

$$\mathrm{Prob}(X_j > \eta | X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta) = \frac{\mathrm{Prob}(X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta)}{\mathrm{Prob}(X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta)}, \quad (27)$$

for $k \geq 2$. Therefore, a practical formula to estimate $\beta_k(\eta)$ is

$$\hat{\beta}_k(\eta) = \frac{\sum_{j=k}^{N} I\big(\{X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\big)}{\sum_{j=k+1}^{N} I\big(\{X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\big)}, \tag{28}$$

where $\sum_{j=k}^{N} I\big(\{X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\big)$ is the counted number of events $\{X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}$, or one exceedance preceded by $k-1$ non-exceedances, and $\sum_{j=k+1}^{N} I\big(\{X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\big)$ is the counted number of non-exceedances or events $\{X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}$, whether they be followed by an exceedance or not. For small values of η, the number of non-exceedances may be 0, making it impossible to calculate $\hat{\beta}_k(\eta)$ for those values. For large η, the number of non-exceedances will be close to N, so that we can approximate

$$\tilde{\beta}_k(\eta) \approx \frac{1}{N} \sum_{kj}^{N} I\big(\{X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\big), \tag{29}$$

which is an alternative estimate for $\beta_k(\eta)$. This formula may often be preferable in practice, since it requires less counting. In the following we will write $\hat{\beta}_k(\eta)$ for the estimates of $\beta_k(\eta)$, but $\tilde{\beta}_k(\eta)$ can of course be substituted.

Now, if we partition our time series into R blocks, and estimate $\beta_k(\eta)$ for each block, obtaining R estimates $\hat{\beta}_k^{(r)}(\eta)$, the mean of those R estimates will be an estimate for the average exceedance rate. The estimate will be

$$\hat{\epsilon}_k(\eta) = \frac{1}{R} \sum_{r=1}^{R} \hat{\beta}_k^{(r)}(\eta). \tag{30}$$

If R is large enough, $R \geq 20$, we can also estimate the standard deviation $s_k(\eta)$ of $\bar{\epsilon}_k(\eta)$,

$$\hat{s}_k(\eta) = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \Big(\hat{\beta}_k^{(r)}(\eta) - \hat{\epsilon}_k(\eta)\Big)^2}, \tag{31}$$

and estimate 95% confidence intervals for $\bar{\epsilon}_k(\eta)$,

$$\hat{\mathrm{CI}}\big(\bar{\epsilon}_k(\eta)\big) = \hat{\epsilon}_k(\eta) \pm \frac{1.96 \hat{s}_k(\eta)}{\sqrt{R}}. \tag{32}$$

Generally, the blocks should not be too short. Each split of the original time series will lessen the number of possible conditional events in the resulting blocks, as we cannot count across the splits. Therefore, two shorter blocks instead of one long will give worse $\bar{\epsilon}_k(\eta)$ estimates for larger k.

The blocks should ideally be of the same length, containing the same number of elements from the time series. It is convenient to have one block per time period. For example, if we have several years of daily observations of a phenomenon, it is convenient to let each year represent a block, let one year be the unit of time, and set $N = 365$.

However, often the number of observations per time unit and the length of the blocks will be different. This may be due to missing data, or because we have chosen to have multiple blocks per period, in order to have more blocks. Thus, it may be the case that the blocks have different lengths $N_1, \ldots, N_R$, or that they have the same length N, but that N is different from the number of observations per period. In such cases, we must be careful. When calculating $\hat{\beta}_k^{(r)}(\eta)$, N should be the length of the particular block in question, but when calculating the return levels using (23), N should be the number of observations per period.

To find an estimate from one realization of a time series, we need to assume that the time series in question is ergodic. However, if this is not the case, we can assume that our time series consists of smaller parts, each of which is ergodic on its own. Making this assumption, it can be argued that our long-time estimates are valid. [9]

The estimations just mentioned are easily made on a computer. In practice, however, the time series will often contain missing observations or NA's. This is handled simply by cutting away the NA's and shortening the time series. If so, the blocks will be of different lengths. Usually this factor is negligible, but if the block lengths are very dissimilar, we might use a weighted mean and standard deviation instead of (30) and (31).

In practice, we must discretize η, making a vector of η points separated by steps of some length d_η. Since the observations will be discretized as well, we should find a discretization that fits the certainty of the data in question. By the nature of (25), choosing a too fine discretization will give us redundant information, while a too coarse discretization will make our estimates less exact than they could be. $\hat{\epsilon}_k(\eta)$ and $\hat{s}_k(\eta)$ will of course be vectors of the same length as η.

2.5 Estimation of the Constants

Having found estimates of $\bar{\epsilon}_k(\eta)$, the next step in the process of estimating the return levels is to find estimates of the constants q, b, a, and c in (22). First, we note that that expression has a much simpler form on the logarithmic scale, where

$$\log \bar{\epsilon}_k(\eta) = f(\eta) = \log q - a(\eta - b)^c, \eta \geq \eta_1. \tag{33}$$

We will call the curve of $\bar{\epsilon}_k(\eta)$ on the logarithmic scale $f(\eta)$, and will refer to a plot of $f(\eta)$ against η as a log plot. In the future, we will mostly work on this scale. Therefore, we will be interested in having the confidence intervals of $\bar{\epsilon}_k(\eta)$ on the same scale. We will call them $\hat{\mathrm{CI}}\big(\log \bar{\epsilon}_k(\eta)\big)$, and approximate them by simply taking the logarithm of the confidence intervals in (32). That is,

$$\hat{\mathrm{CI}}\big(\log \bar{\epsilon}_k(\eta)\big) \approx \log \left\{ \hat{\epsilon}_k(\eta) \pm \frac{1.96 \hat{s}_k(\eta)}{\sqrt{R}} \right\}. \tag{34}$$

This approximation will give us problems when $\frac{1.96\hat{s}_k(\eta)}{\sqrt{R}} < \hat{\epsilon}_k(\eta)$ or $\frac{1.96\hat{s}_k(\eta)}{\sqrt{R}} \approx \hat{\epsilon}_k(\eta)$, but as this will only happen for very large values of η, the approximation can probably be used for most η values.

To estimate the constants q, b, a, and c, we are going to fit a curve to the estimates $\hat{\epsilon}_k(\eta)$, for $\eta \geq \eta_1$. The fitted curve will be

$$\hat{f}(\eta) = \log \hat{q} - \hat{a}(\eta - \hat{b})^{\hat{c}}, \eta \geq \eta_1, \tag{35}$$

and $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ will be our estimates of the constants q, b, a, and c, respectively. So, our problem turns out to be a curve fitting problem. However, fitting a non-linear curve determined by four constants to $\log \hat{\epsilon}_k(\eta)$ will be difficult, unless we can find more information to help us.

Now, we take the logarithm on both sides of (33). It then turns out that

$$\begin{aligned} \left|\log\left[\frac{\bar{\epsilon}_k(\eta)}{q}\right]\right| &= a(\eta - b)^c \\ \log\left|\log\left[\frac{\bar{\epsilon}_k(\eta)}{q}\right]\right| &= \log a + c\log(\eta - b), \end{aligned} \tag{36}$$

for $\eta \geq \eta_1$. This implies that $\log\left|\log\left[\frac{\bar{\epsilon}_k(\eta)}{q}\right]\right|$ is linear with respect to $\log(\eta - b)$. So, a plot of $\log\left|\log\left[\frac{\bar{\epsilon}_k(\eta)}{q}\right]\right|$ against $\log(\eta - b)$ will give a straight line for $\eta \geq \eta_1$. Such a plot will be referred to as a log log-log plot.

Accordingly, good estimates $\hat{q}$ and $\hat{b}$ of q and b should give us a straight line when plotting $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{\hat{q}}\right]\right|$ against $\log(\eta - \hat{b})$. This information is useful in three ways. First, it gives us a criterion for finding good estimates $\hat{q}$ and $\hat{b}$, which are the ones that give us a straight line in the log log-log plot. Second, it gives us a criterion for finding η_1, since η_1 should be the value of η where $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{\hat{q}}\right]\right|$ becomes straight in the log log-log plot. And third, we can then easily find estimates of the constants a and c. Having found good estimates $\hat{q}$ and $\hat{b}$, the last two estimates $\hat{a}$ and $\hat{c}$ can be found simply by doing a linear fit on the straight line and calculating the slope and the intercept.

Obviously, we could complete the estimation of the constants in the log log-log plot. However, the estimates will be much more exact if we do the curve fitting in the log plot. Therefore, the results obtained by studying the log log-log plot can be regarded as initial estimates to be used as start points for the analysis in the log plot. Having good start points, the curve fitting of (35) will be much easier.

Hence, our analysis will be a two-step operation. In step 1, we fit a straight line to $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{\hat{q}}\right]\right|$ in the log log-log plot, and in step 2 we fit a non-linear curve to $\log \hat{\epsilon}_k(\eta)$ in the log plot. The resulting estimates $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ enable us to estimate the m periods return levels by (23).

When performing the analysis, we cannot use all the available estimates $\hat{\epsilon}_k(\eta)$. First, we can only use $\hat{\epsilon}_k(\eta)$ for $\eta \geq \eta_1$, because the approximation (22) is only valid for $\eta \geq \eta_1$. Second, for large values of η, we will have few data with which to estimate $\bar{\epsilon}_k(\eta)$, and the resulting estimates $\hat{\epsilon}_k(\eta)$ will be unreliable. These unreliable estimates might make the curve fitting difficult. Therefore, we should ignore the estimates $\hat{\epsilon}_k(\eta)$ where $\eta > \eta_2$, for some η_2, and only take into account those estimates where $\eta_1 \leq \eta \leq \eta_2$.

So, in addition to the four constants q, b, a and c, which must be estimated, we will have two more constants, η_1 and η_2, that we must decide on. Obviously, it would be difficult to find all six constants at once. As mentioned, we will instead use a two step method. The method will be sketched here, and the details will be given in the next two sections.

Step 1

- For all possible combinations of η_1 and η_2:
  - For all possible combinations of q and b:
    - Try to fit a straight line to the estimates in the log log-log plot.
- Select the best log log-log plot fit, choose the corresponding η_1 and η_2, and let the corresponding estimated constants $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ be starting values in the next step.

Step 2

- Using the starting values found in step 1, fit a curve to the estimates in the log plot. The resulting estimated constants $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ will be our final estimates, with which we can estimate the return levels.

In general, we should try to gather as much information as possible about the data to be analysed. Especially, finding bounds on the possible values of the constants will make the analysis easier.

As we proceed to demonstrate a practical method for estimating the return levels, we note that in practice, we operate with six different vectors of the same length. As mentioned, we have a η vector, having been chosen by us as a discretization of the η scale. In addition, there are corresponding vectors $\hat{s}_k(\eta)$, $\hat{\epsilon}_k(\eta)$, and $\log \hat{\epsilon}_k(\eta)$, as well as two vectors for the lower and the upper confidence bounds of $\log \bar{\epsilon}_k(\eta)$.

2.6 Step 1

As noted above, the purpose of step 1 is to find a good fit, a straight line, in the log log-log plot. We therefore need a criterion by which to judge our linear fits, so as to decide which is a "good" fit. First, we obviously want to use as much of our vector $\hat{\epsilon}_k(\eta)$ as possible, as the fit will be more precise the more points we use. In other words, we want η_1 to be as small as possible, and η_2 as large as possible, while still being able to have a good fit. By keeping η_1 as small as possible, we also ensure that we use the most reliable estimates of $\bar{\epsilon}_k(\eta)$. [13] So, we introduce the distance

$$L = \eta_2 - \eta_1, \tag{37}$$

which we want to maximize.

Second, we need a measure to use as a criterion for deciding how well the linear fit suits the data. Since we must compare fits based on different sets of $\hat{\epsilon}_k(\eta)$ points, we

cannot simply use least squares minimization. The fitted line in the log log-log plot will be of the form

$$y = \lambda + \kappa \log(\eta - b), \tag{38}$$

where κ is the slope, and λ is the intercept. Now, the linearity of the fit can be measured by the distance

$$\Delta = \max_{\eta_1 \leq \eta \leq \eta_2} \left| \log\left| \log\left[\frac{\hat{\epsilon}_k(\eta)}{q} \right] \right| - y \right|, \tag{39}$$

which is the maximum distance between the estimates $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|$ and the fit y for values of η between η_1 and η_2. As will be discussed, this is not a perfect measure, but it probably is good enough for our use. Obviously, the better the linear fit, the smaller Δ will be.

Hence, we have two criteria for deciding on what is a good fit. First, L should be as large as possible, and second, Δ should be as small as possible. Clearly, these two criteria will not always be compatible. We will have to find a compromise, so as to render justice to both.

Now, having found criteria by which to judge our linear fits, we turn to the practical procedure. First, we will try to decide on which η_1 and η_2 to use. Usually, we do not want to search through all the possible values of η when looking for the most suitable values of η_1 and η_2. If the discretization of the η scale is fine, there can be hundreds of such combinations, and investigating all of them would be a time-consuming task. But it is clear that the return levels cannot be overly sensitive to the choice of η_1 and η_2. If we assume that the approximation in (22) is valid for η values larger than some optimal η_1 value $\eta_1^{(opt)}$, choosing a slightly larger η_1 value will still make the approximation valid. Hence, we can allow a slight uncertainty in the choice of η_1. The same will be true for η_2.

Likely values of η_1 and η_2 can often be read from a plot of $\log \hat{\epsilon}_k(\eta)$ against η. Using this information, we can make selections of some of the more likely η_1 and η_2 values. On the η scale, we select a part $(\eta_1^{(min)}, \leq \eta_1^{(max)})$ which we believe contains η_1. This selection will correspond to a certain number of elements of the η vector. Now, if the discretization of η is coarse, we may use all those elements. If not, to save computer run time, we choose only n_{η_1} of the elements to represent the entire selection $(\eta_1^{(min)}, \eta_1^{(max)})$. For example, if there are really 100 elements contained in the selection, we may choose only 10 of them as our representation. These representative elements should of course be dispersed equally throughout the selection. The selected n_{η_1} elements from the η vector will be a vector of possible η_1 values. Equivalently, we make a selection $\eta_2^{(min)} \leq \eta \leq \eta_2^{(max)}$ which we believe will contain η_2, and choose n_{η_2} so as to get a vector of possible η_2 values.

The width of the η_1 and η_2 selections depends on our knowledge about η_1 and η_2. If it is fairly evident where η_2, must be, we can let it be small, while if we are uncertain, we must leave it wider. We can obtain such knowledge by looking at the plot of our estimates $\log \hat{\epsilon}_k(\eta)$ along with the corresponding confidence intervals. η_1 will be found where the

curve of $\hat{\epsilon}_k(\eta)$ estimates straightens out, while η_2 will be found where the confidence intervals broaden and the estimates $\hat{\epsilon}_k(\eta)$ are seen to become irregular compared to the estimates for lower η. The larger the k values, the fewer conditional exceedances we will have for smaller η values, and the more uncertain the estimates of $\log \epsilon_k(\eta)$ will be for those values. Hence, we if we have already found η_1 for $k = 1$, we should expect η_1 to be larger for $k = 2$, and so on. This information may sometimes be useful.

All in all, we will now have $n_{\eta_1} \cdot n_{\eta_2}$ possible combinations of η_1 and η_2 values. For each of these combinations, we must try to find the best linear fit in the log log-log plot.Hence, we need to find the q and b values that will give us the best linear fit. This we will do simply by running through all the possible values of q and b, and try to fit a linear curve in the corresponding log log-log plot. Obviously, we must discretize the possible q and b values, using steps d_q and d_b, respectively. This discretization will give us a vector of length n_q of possible values of q and a vector of length n_b of possible b values. Experience shows that the value of q is much more important for finding a linear fit than the value of b. Therefore, q should have a finer discretization than b.

What values can q and b possibly take? From (36) we see that we must have $q > \max_{\eta_1 \leq \eta \leq \eta_2} \hat{\epsilon}_k(\eta)$ and $b < \min_{\eta_1 \leq \eta \leq \eta_2} \eta = \eta_1$, since $\eta - b$ must be positive and $\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]$ must be of the same sign for all $\eta_1 \leq \eta \leq \eta_2$. We thus have a lower and an upper bound for q and b, respectively. Obviously, these will vary with each combination of η_1 and η_2. We must select some upper bound $q^{(high)}$ for q and some lower bound $b^{(low)}$ for b ourselves. $q^{(high)}$ we must select entirely from experience, but $b^{(low)}$ may often be inferred from the nature of the physical phenomenon we are studying. For example, in some cases it turns out that $b^{(low)} = 0$.

Now, for each combination of η_1 and η_2 values, we run through all combinations of the constants q and b, making linear fits of $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|$ against $\log(\eta - b)$. Comparing (36) against (38), we can calculate the corresponding constants a and c from the slope and intercept of the linear fit by

$$a = \exp(\kappa) \tag{40}$$

and

$$c = \lambda. \tag{41}$$

We also calculate the Δ value for the particular fit from (39).

So, for each combination of η_1 and η_2 values, we will have $n_q \cdot n_b$ different Δ values, each of which corresponds to a set of constants q, b, a and c of a particular linear fit. It is also useful to compute a return level corresponding to the fit, for exemple the 100 years return level, for comparison. To each combination of η_1 and η_2, we assign the lowest of these Δ value along with its corresponding constants and return level. We then go on to compare the Δ values of combinations of η_1 and η_2 values that have the same L value. To each L value we assign the lowest Δ value together with its corresponding η_1, η_2 values, constants and return level. Hence, each value of L will have a corresponding

return level and a Δ value, which represents a particular fit, which is itself represented by η_1, η_2, q, b, a, and c.

This part of the analysis will of course be done by a computer, which will present us with a table of L values and the corresponding Δ, η_1, η_2, q, b, a, c, and x_m values. However, we must do the last part of the analysis ourselves. This analysis consists in comparing the fits of the different L values and choosing the one which appears to be the best. In doing so, we must consider our two criteria. Each fit has a corresponding pair of L and Δ values. As noticed, we want L to be as large as possible, while Δ should be as small as possible. The ideal fit is therefore a fit which has both a large value of L and a reasonably small value of Δ. For example, if several fits have very similar Δ values, we select the one with the largest L value. At last, we will have found one best fit with corresponding η_1 and η_2 values among the possible choices in the table. The constants of this fit will be our initial estimates $\hat{q}_i$, $\hat{b}_i$, $\hat{a}_i$, and $\hat{c}_i$.

Choosing the best fit from the values of L and Δ only may be difficult. It is therefore advisable to make both log log-log plots and log plots of the estimates and the fitted curves. In these plots, the fits should of course follow the estimates as closely as possible for $\eta_1 \leq \eta \leq \eta_2$. Generally, the log plots are more useful for comparison between the fits than the log log-log plots.

Further, the return level estimates of each fit can be useful for comparison between the fits. As mentioned, for η_1 values that are larger than the optimal η_1 value, (22) will still be valid. Hence, the return level estimates made with η_1 values larger than the optimal value should be approximatively the same. If we plot the return levels against the η_1 values, the optimal η_1 value can then be found where the return levels seem to stabilize. A plot of the return levels against the η_2 values should give a similar result, as a η_2 value smaller than the optimal η_2 should give approximatively the same return levels. A plot of the return levels against the L may also be able to give some information.

Combining the information obtained from these sources, we should be able to decide which of the fits is the best. The most important is of course that the fit actually is a good fit in the log plot, since we are going to operate on the log scale in the next step of the analysis. On the other hand, we should not put too much emphasis on this step, since the results are only starting points for step 2. Only η_1 and η_2 are actually kept, while the estimates of q, b, a, and c will be altered. We must also remember that a certain uncertainty is allowed in the choice of η_1 and η_2. Generally, we should try to make step 1 as simple as possible, by not choosing n_{η_1} and n_{η_2} too large and d_q and d_b too small. If not, the analysis will be impractical and computer run time too long.

Especially, the analysis will become complicated and require much computer run time if n_{η_1} and n_{η_2} are too large. The number of fits to compare in the end will simply be too great. In practice, n_{η_1} and n_{η_2} should be no more than 10 or 15. But if we are uncertain about where to find η_1 and η_2, we may require a broad selection of possible values of η_1 and η_2. Selecting only 10 possible values of η_1 will then give very inaccurate results. However, after the analysis, we can perform it again, but this time with a tighter selection of η_1 and η_2. Each performance of the analysis will give us more accurate information on where to find the optimal values of η_1 and η_2. Thus, we can "zoom in" on the optimal

values. This procedure is usually more effective than one analysis with large n_{η_1} and n_{η_2} values. The same argument holds true for small d_q and d_b values.

Finally, we must add a few comments about the measure Δ, which is of crucial importance in the analysis. Since the fits that are compared using Δ have different q and b values, and we measure the goodness of these fits in the log log-log plot, where the scales are dependent on those two constants, our measure is not an absolute measure, since a distance in one log log-log plot cannot be compared directly to another distance in another log log-log plot. But by transforming those distances to some common measure, we must leave the log log-log plot, and loose what we were going to measure in the first place, the linearity. Therefore, we choose to keep our measure, knowing it to be imperfect. On the other hand, if we suppose that there exists some optimal value $\Delta^{(opt)}$ for each value of L, having corresponding constants $q^{(opt)}$ and $b^{(opt)}$, then, in an area around $q^{(opt)}$ and $b^{(opt)}$, we suppose the Δ values to be relatively stable, justifying comparisons in that area.

2.7 The Numerics

We will not go into details on the implementation of step 1 on a computer. Instead, the R code of an actual implementation is given in Appendix A. However, when performing the computations on a computer, we may save some run time by using a special procedure.

As we have seen, for each combination of η_1 and η_2 values, and for each combination of possible q and b values, we make a linear fit. Each fit takes a certain amount of computer run time. Obviously, since only certain q and b values give good fits, most of those fits will be poor. Hence, since we are only interested in the good fits, we can save run time by avoiding the calculation of the poorer fits.

We can do this by comparing the fit under consideration to the best Δ value of those fits which have already been performed, and which have the same value of L. Let us call the current best Δ value Δ^*. In order that the values q and b under consideration may give a better fit than that represented by Δ^*, from the definition (39) of Δ, all the points $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|$ have to be within a distance of Δ^* from the linear fit in the log log-log plot. Equivalently, this line must be within a distance of Δ^* from any of the points $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|$, and also, more specially, from the first and the last among these points, the points $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|^{(first)}$ and $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|^{(last)}$.

Now, if we draw a line between $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|^{(first)}$ and $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|^{(last)}$, and draw two parallel lines a distance Δ^* away from that line on both sides, our fit has to be found between those two lines if it is to be better than the Δ^* fit. Further, if we draw two new parallel lines $2\Delta^*$ away from the first line on both sides, all the points $\hat{\epsilon}_k(\eta)$ have to be found between those two lines if we are going to get a better fit. If not, there is no use of making the fit. By performing this test, we will be able to avoid some of the calculations. The concept is illustrated in figure 1. We may also be able to save som computer run time by starting with the largest η_1 values, since it will be easier to make

a linear fit for η_1 values larger than the hypothetical optimal value $\eta_1^{(opt)}$, than for η_1 values smaller than that value. Hence, we will obtain small Δ values from the start, and be able to reject many of the later fits at once.

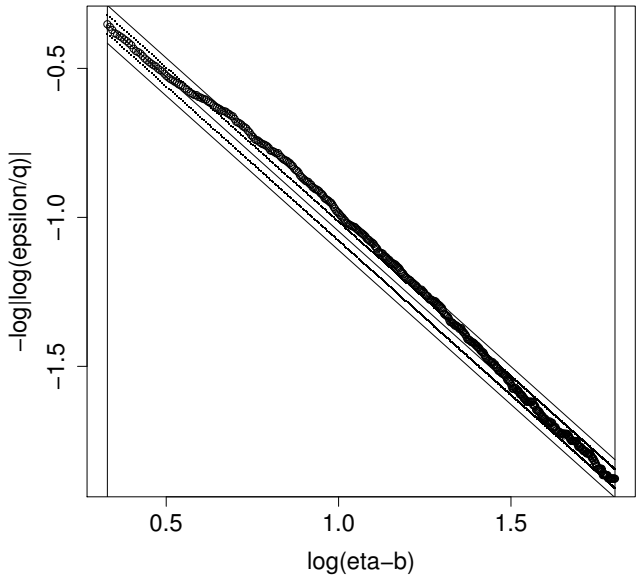


Figure 1: Illustration of the principle used to save computing time. All the points $\log\left|\log\left[\frac{\hat{\epsilon}_k(\eta)}{q}\right]\right|$ have to be situated between the outer lines if this particular combination of q and b can possibly give us a better fit than we already have. Hence, this particular combination of q and b could be rejected at once.

2.8 Step 2

In step 1 we found a good linear fit in the log log-log plot. This fit had corresponding η_1 and η_2 values, and gave us initial estimates $\hat{q}_i$, $\hat{b}_i$, $\hat{a}_i$ and $\hat{c}_i$ of the constants q, b, a, and c. Now, we keep the η_1 and η_2 values, while we use the initial estimates of the constants as starting values for a further curve fitting in the log plot.

The initial estimates were found using the linear property of the log log-log plot. When finding the final estimates, we will use the log plot. This step can be seen as refining the initial estimates, since we expect that a fit in the log plot will give a higher degree of accuracy than the linear fit in the log log-log plot. As we have seen, in the log plot we need to fit a curve $\hat{f}(\eta)$ given by (35) to the points $\hat{\epsilon}_k(\eta)$, for $\eta_1 \leq \eta \leq \eta_2$. This is a nonlinear problem, but it can be solved numerically using a nonlinear least squares solver.

Since our estimates $\hat{\epsilon}_k(\eta)$ have varying certainty, we should probably put more weight

on the more certain estimates when fitting the curve, while less emphasis should be put on the less certain estimates. Therefore, we will use a weighted version of the nonlinear least squares solver. To each of the elements in the vector of estimates $\hat{\epsilon}_k(\eta)$, we assign a weight $w(\eta)$.

Now, our curve fitting problem turns out to be equivalent to finding the constants $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ that minimize the expression

$$\sum_{\eta_1 \leq \eta \leq \eta_2} \Big(\log \hat{\epsilon}_k(\eta) - f(\eta)\Big)^2 w(\eta) = \sum_{\eta_1 \leq \eta \leq \eta_2} \Big(\log \hat{\epsilon}_k(\eta) - \log q + a(\eta - b)^c\Big)^2 w(\eta), \quad (42)$$

where η and $\log \hat{\epsilon}_k(\eta)$ are vectors, and where $w(\eta)$ is a vector of weights. We will return to the choice of these weights further below. η_1 and η_2 are of course the η_1 and η_2 values we found in step 1.

The weighted nonlinear least squares problem can be solved numerically by the Gauss-Newton or the Marquardt-Levenberg algorithm, for example using the *lsqnonlin* method in Matlab [6, p. 11-178]. We then let the values of q, b, a and c vary, but we should not let them vary freely. Especially, either q or b should be have both lower and upper bounds, while a and c should be positive. For q and b, we can use the same bounds as we used when finding the linear fits in step 1, that is $\big(\max_{\eta_1 \leq \eta \leq \eta_2} \hat{\epsilon}_k(\eta), q^{(high)}\big)$ for q and $\big(b^{(low)}, \min_{\eta_1 \leq \eta \leq \eta_2} \eta = \eta_1\big)$ for b. However, only one of them needs to be bounded on both sides, since they are related. If $b^{(low)}$ is determined by the nature of the data under study, it is convenient to let q vary freely upwards, since $q^{(high)}$ was only an arbitrary number chosen in step 1. Experience shows that if q and b are not bounded at all, we will often get estimates $\hat{q}$ and $\hat{b}$ that tend to ∞ and $-\infty$, respectively.

As starting values to the numerical method we use the initial estimates which we found in step 1, $\hat{q}_i$, $\hat{b}_i$, $\hat{a}_i$, and $\hat{c}_i$.

In some cases, it may be difficult to reach an optimal solution. As we saw, the case $c = 1$ corresponds to the Gumbel distribution. From (22) it is clear that if $c = 1$ there will be an infinity of possible combinations of q and b that can describe the same curve $f(\eta)$. Thus, if $c \approx 1$, the numerical method may have problems finding the optimal fit. If we experience problems, it may perhaps be better to set $\hat{c} = 1$ at once, fix either q or b, and run the curve fitting varying only the two remaining variables. [11]

The result of the nonlinear least squares fit will be a fitted curve $\hat{f}(\eta)$, represented by the estimates $\hat{q}$, $\hat{b}$, $\hat{a}$ and $\hat{c}$, from which we can calculate the return level estimates $\hat{x}_m$ by inserting $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$ for q, b, a, and c in (23). This procedure can be regarded as extrapolating our curve $\hat{f}(\eta)$ in the log plot outside of $\eta_1 \leq \eta \leq \eta_2$, and finding x_m as the value of η where $\hat{f}(\eta)$ crosses the level on the y axis of the plot corresponding to $\bar{\epsilon}_k(x_m)$. Thus, it is important that the phenomenon being studied is described by a homogeneous distribution. If it is not, and for example the larger values belong to a different distribution than the smaller, the extrapolation will be invalid, and the AER method is rendered useless.

2.9 Weights

We have to decide upon a set of weights to use when fitting the nonlinear curve. The purpose of using weights is to put less emphasis on the more uncertain estimates $\log \hat{\epsilon}_k(\eta)$. As we are now operating on the logarithmic scale, the uncertainty is expressed by the width of the confidence intervals of $\log \bar{\epsilon}_k(\eta)$ given in (34). To put emphasis on the more certain estimates, we can use as weights the inverse of the confidence interval width on the log scale. The weights formula will then be

$$w_1(\eta) = \frac{1}{\hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))^{(high)} - \hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))^{(low)}}. \tag{43}$$

Of course, for such η where $\hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))$ is undefined, we set $w_1(\eta) = 0$. Alternatively, we can use the inverse of the squared confidence interval width,

$$w_2(\eta) = \frac{1}{\left(\hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))^{(high)} - \hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))^{(low)}\right)^2}. \tag{44}$$

Both weights formulae can be used, but they will give slightly different results, as the second formula puts much more emphasis on the more certain estimates of $\log \hat{\epsilon}_k(\eta)$, while almost neglecting the more uncertain. Now, the more uncertain estimates will be found where η is great, in the tail of the distribution. It is of course a good thing that emphasis is put on the reliable estimates, but as we are primarily interested in tail behaviour of $\bar{\epsilon}_k(\eta)$, we should not neglect the information contained in $\log \hat{\epsilon}_k(\eta)$ about the tail. Therefore, the weights formula $w_1(\eta)$ probably is the better choice.

2.10 Confidence Intervals

Having found estimates $\hat{x}_m$ for the return levels x_m, we want to estimate confidence intervals for x_m.

A straightforward way of finding those confidence intervals is to use the confidence intervals of $\log \bar{\epsilon}_k(\eta)$. The uncertainty of the estimates $\log \hat{\epsilon}_k(\eta)$ is described by the confidence intervals $\hat{\mathrm{CI}}(\log \bar{\epsilon}_k(\eta))$ of (34), which give us confidence bands on each side of $\log \hat{\epsilon}_k(\eta)$ for $\eta_1 \leq \eta \leq \eta_2$. We realize that the fitted curve $\hat{f}(\eta)$ is only one of the possible fitted curves defined by (33), as we could think of other curves defined by other constants $\hat{q}$, $\hat{b}$, $\hat{a}$, and $\hat{c}$. Each of those curves would provide a particular estimate $\hat{x}_m$ of the return level. Now, we assume that the return levels corresponding to those fitted curves that fall inside of the 95% confidence bands around $\hat{\epsilon}_k(\eta)$ will describe a 95% confidence interval for the return level x_m. Then the minimum and maximum values of those return level estimates would constitute the lower and upper confidence interval bounds, respectively, for x_m. [11]

So, if we want to estimate a confidence interval for x_m, we should search for the two curves $\hat{f}^{(high)}(\eta)$ and $\hat{f}^{(low)}(\eta)$ that fall inside the confidence bands on both sides of $\log \hat{\epsilon}_k(\eta)$ for $\eta_1 \leq \eta \leq \eta_2$ and at the same time give the highest and lowest possible return levels estimates, $\hat{x}_m^{(high)}$ and $\hat{x}_m^{(low)}$, respectively. Then $\hat{x}_m^{(high)}$ and $\hat{x}_m^{(low)}$ constitute the

boundaries of the estimated confidence interval for the return level x_m. Each of the two curves $\hat{f}^{(high)}(\eta)$ and $\hat{f}^{(low)}(\eta)$ will be defined by constants $\hat{q}^{(high)}$, $\hat{b}^{(high)}$, $\hat{a}^{(high)}$, $\hat{c}^{(high)}$, and $\hat{q}^{(low)}$, $\hat{b}^{(low)}$, $\hat{a}^{(low)}$, $\hat{c}^{(low)}$, respectively.

However, since the edges of the confidence bands on both sides of $\log \hat{\epsilon}_k(\eta)$ will be rugged, the confidence interval estimates obtained in this way will probably be too short. We can ameliorate the situation by transferring the confidence intervals of $\log \bar{\epsilon}_k(\eta)$ from the estimates $\log \hat{\epsilon}_k(\eta)$ onto the fitted curve $\hat{f}(\eta)$, which is smooth. Then, the confidence bands will be smoother, and the resulting estimated confidence intervals of x_m broader. Hence, for a 95% confidence interval, we approximate

$$\hat{\mathrm{CI}}\big(f(\eta)\big) \approx \log\Big[\hat{q}\exp\{-\hat{a}(\eta-\hat{b})^{\hat{c}}\} \pm \frac{1.96\hat{s}_k(\eta)}{\sqrt{R}}\Big], \eta_1 \le \eta \le \eta_2. \tag{45}$$

This will give us smooth bands with which to find the curves that represent the confidence intervals boundaries for x_m. However, it should be stressed that the curves $\hat{f}(\eta)^{(high)}$ and $\hat{f}(\eta)^{(low)}$ themselves do not constitute any confidence intervals; only the corresponding estimates $\hat{x}_m^{(high)}$ and $\hat{x}_m^{(low)}$ of x_m do.

The formula in (45) should be used with some care, though. As noted earlier, when the difference $\hat{q}\exp\{-\hat{a}(\eta-\hat{b})^{\hat{c}}\} - \frac{1.96\hat{s}_k(\eta)}{\sqrt{R}}$ approaches 0, which will sometimes be the case for the lower confidence bound for large values of η, the resulting confidence bands will be unreliable and may be rugged, such that the confidence intervals of x_m will become tighter than they should. And when that difference is negative, we will have no bounds at all. One should therefore consider the confidence intervals calculated in this way before using them, and, if necessary, cut away the parts that are seen to present a problem.

The problem of finding the two curves $\hat{f}^{(high)}(\eta)$ and $\hat{f}^{(low)}(\eta)$ can be regarded as an optimization problem. We want to find the values $\hat{x}_m^{(low)}$ and $\hat{x}_m^{(high)}$ that are as small and as large as possible, respectively, under the condition that the curves, which are determined by the corresponding constants $\hat{q}^{(low)}$, $\hat{b}^{(low)}$, $\hat{a}^{(low)}$, $\hat{c}^{(low)}$, and $\hat{q}^{(high)}$, $\hat{b}^{(high)}$, $\hat{a}^{(high)}$, $\hat{c}^{(high)}$, respectively, do not fall outside of the confidence bands $\hat{\mathrm{CI}}\big(f(\eta)\big)$, for $\eta_1 \le \eta\eta_2$. This is a constrained nonlinear optimization problem, which can be solved numerically, for example by using the *fmincon* method in Matlab [6, p. 11-35]. We let all the four constants vary, but here, too, they cannot be unbounded, especially not q and b. We can use the same bounds as we used with the curve fitting in section 2.8.

Generally, however, we should use all that we know about our data to find tighter bounds on the constants and thus keep the confidence intervals from becoming too wide. Especially, having bounds on c would be useful. The further away from 1 c is allowed to be, the more rounded the curves will be, and consequently, the wider the confidence intervals of x_m will be. Again, it is also important that η_1 is as small as possible, as the confidence bands will be tighter for smaller values of η.

3 The Data

We are going to test the AER method on four different data sets. The first two are two series of wind speed observations, while the two latter are two series of ocean wave height observations.

3.1 The Wind Speed Data

Our first data sets consist of two data series of 20 years of daily wind speed observations from the weather stations of the Norwegian Meteorological Institute at Ørlandet Airport and Alta Airport in Norway.

The Norwegian Meteorological Institute makes daily observations of many different meteorological phenomena at their weather stations. Especially, they make certain main observations at specific times. One of the phenomena which are measured in this way is the wind speed. The main observation of the wind speed is called FF, and is defined to be the average of the wind speed during the last ten minutes before the observation time. On day number i we will thus have n wind speed measurements $\{V_{i1}, \ldots, V_{in}\}$ from the n main observations. The largest among these measurements,

$$X_i = \max_{j=1,\ldots,n} V_{ij}, \tag{46}$$

is called FFX. [7]

We will investigate a series of such FFX observations from the weather station Ørland III at Ørlandet Airport in Sør Trøndelag fylke in Norway, and from the weather station at Alta Airport. Ørlandet is situated on a flat plain at the western tip of the Fosen peninsula, by the Norwegian Sea. Given the location, we should expect to find some high wind speeds among the observations. Alta is located in the extreme north of Norway, but as it is located by the innermost part of a fjord, we would expect the wind speeds to be lower.

Our data series consists of all FFX measurements at Ørlandet and at Alta Airports from January 1, 1987, to December 31, 2006. All in all, for each location we have 7305 measurements from a period of 20 years. The numbers are given in meters per second, $[\frac{m}{s}]$, and the vector of observations is called X,

$$X = \{X_1, \ldots, X_{7305}\}. \tag{47}$$

The observations were made by the Norwegian Meteorological Institute and were collected for our purpose using their web service *eKlima* [8]. Regarding the quality of the material, the Meteorological Institute themselves regard them as being "a little uncertain" (*litt usikre*). For our purpose, however, the reliability of the data should not be of too great concern, as we are primarily using the data to test the AER method. A few observations are missing, but since we have a large number of observations, this should be insignificant.

By studying the numbers of the data series, we discover that the observed wind speed values are quite coarsely discretized. The observation values are only to be found

on certain levels. For example, at Ørlandet, for wind speeds between 5 and 10, an observation X_i can only take values such as

$$\{5.1, 5.7, 6.2, 6.7, 7.0, 7.2, 7.7, 8.2, 8.7, 9.3, 9.8\}. \tag{48}$$

As we see, the difference between the possible values of X_i varies slightly, but it is at most 0.5. This is a rather coarse discretization, and these numbers tell us something about the uncertainty in the data material. At Alta the discretization is even coarser, with differences between the layers of possible values at most 0.6, expect for the last year of the series, where it is 0.1. The two parts of the Alta data series thus have different accuracies, but this fact will have no consequence for our results, as we must adjust to the data with the least accuracy.

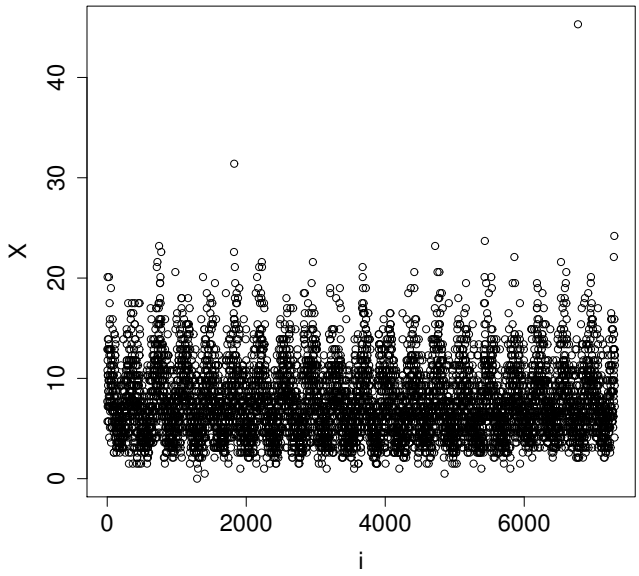


Figure 2: Plot of the Ørlandet wind speed data series.

In figure 2 and 3, we see plots of the entire series X of the Ørlandet and the Alta data, respectively. For the Ørlandet data, most of the values fall between 0 and 25, but we also have two extraordinarily high values, at about 30 and 45, respectively. We may suspect these two values are outliers that do not belong to the distribution we wish to investigate, and as such they might cause problems in a POT analysis. However, since we are going to use the AER method, we do not need to think about that, since our method ignores the data from the very tail $\eta > \eta_2$, where uncertainty is high. For the Alta data, the observed wind speeds are seen to be somewhat lower, as most values fall between 0 and 20. In this plot, wee clearly see how the observed values are found only

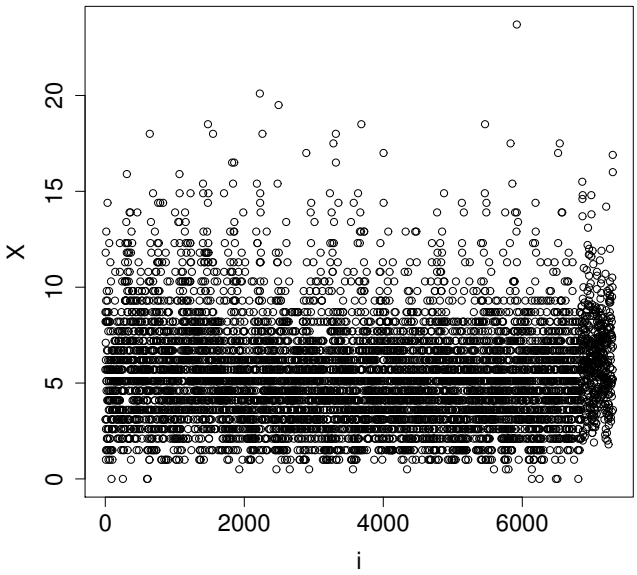


Figure 3: Plot of the Alta wind speed data series.

year	1	2	3	4	5	6	7	8	9	10
max. wind speed	20.1	21.6	23.2	20.1	22.6	31.4	21.6	18.5	21.6	19.0
year	11	12	13	14	15	16	17	18	19	20
max. wind speed	21.1	19.0	23.2	20.6	23.7	20.1	22.1	21.6	45.3	24.2

Table 1: The yearly observed maxima of the Ørlandet wind speed data.

on certain discrete levels, as they seem to fall along horizontal lines. This is not the case for the last year of the data series, where the discretization is finer.

In tables 1 and 2 we list the maximum wind speeds for each of the 20 years, for the Ørlandet and the Alta data, respectively. We again notice the extraordinary values in the 6th and in the 19th year of the Ørlandet data.

Each of the data series has a clear periodicity; this will be clearer if we look at a plot of the data series for the first year only of the Ørlandet data, as given in figure 4. We here see a clear tendency, the highest values occurring at the beginning and at the end of the year, which is quite natural, as we would expect higher wind speeds during the autumn and winter storms. Further, from our knowledge of meteorology we expect that observations made at small time intervals will be correlated, as they might come from the same storm or the same weather system. We will have to account for this correlation in our analysis.

To use the AER method, we assume that the distribution of wind speed maxima

year	1	2	3	4	5	6	7	8	9	10
max. wind speed	15.9	18.0	15.9	15.4	18.5	15.4	20.1	17.0	14.9	14.9
year	11	12	13	14	15	16	17	18	19	20
max. wind speed	17.0	12.9	12.9	14.9	18.5	17.5	23.7	17.5	15.5	14.8

Table 2: The yearly observed maxima of the Alta wind speed data.

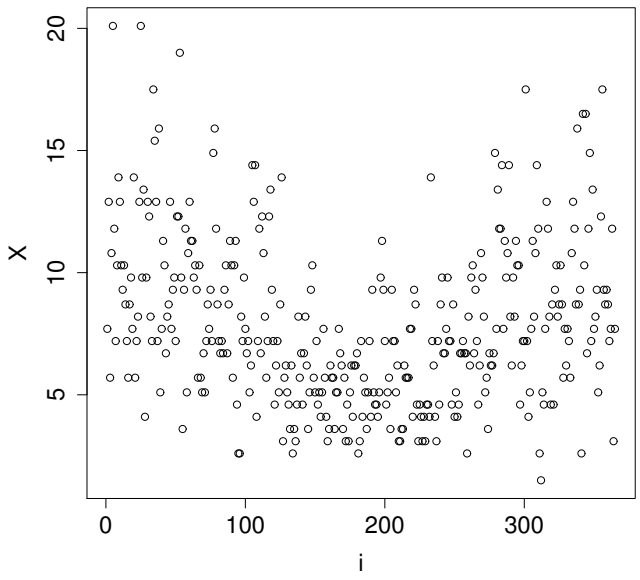


Figure 4: Plot of the first year of the Ørlandet wind speed data.

asymptotically will follow a Gumbel distribution. This seems reasonable, since using a Weibull distribution would implicate that there were some fixed upper limit of the wind speeds only slightly above the maximum observed values. However, there are no physical phenomenon which would seem to explain the existence of such a fixed limit. Further, by looking at the long term distribution of the wind speeds as estimated from observations, it seems clear that the constant b of the Gumbel distribution must be nonnegative, as it is related to the mean wind speed. And as that constant is the same constant as appears in (22), we can assume beforehand that $b \geq 0$. [10]

3.2 The Ocean Wave Data

Our second pair of data series consists two series of 12 years and 8 months of daily ocean wave height observations, also from the Norwegian Meteorological Institute [1].

Wave heights can be defined by the zero up crossing method. We have a level called 0. The wave height H_i of observation number i is defined as the distance between the highest and the lowest sea level during the period T_i between two up crossings or down crossings of the 0 level.

We will use three different types of wave data. The first is called the significant wave height, which is abbreviated H_S. The significant wave height is defined as the average of the $\frac{1}{3}$ largest wave height observations in the observation period [4, ch. 3]. The two other types are called wind sea, abbreviated H_{SWS}, and swell, abbreviated H_{SSW}, respectively. Both of these are components of the significant wave heights. The wind sea is the component which accounts for the part of the wave height which is due to wind, while the swell component accounts for the part of the wave height which is due to far-travelling waves generated in other areas of the sea [5, ch. 5]. Accordingly, the values of H_{SWS} and H_{SSW} must be smaller than the corresponding value of H_S, since the two former constitute parts of the latter.

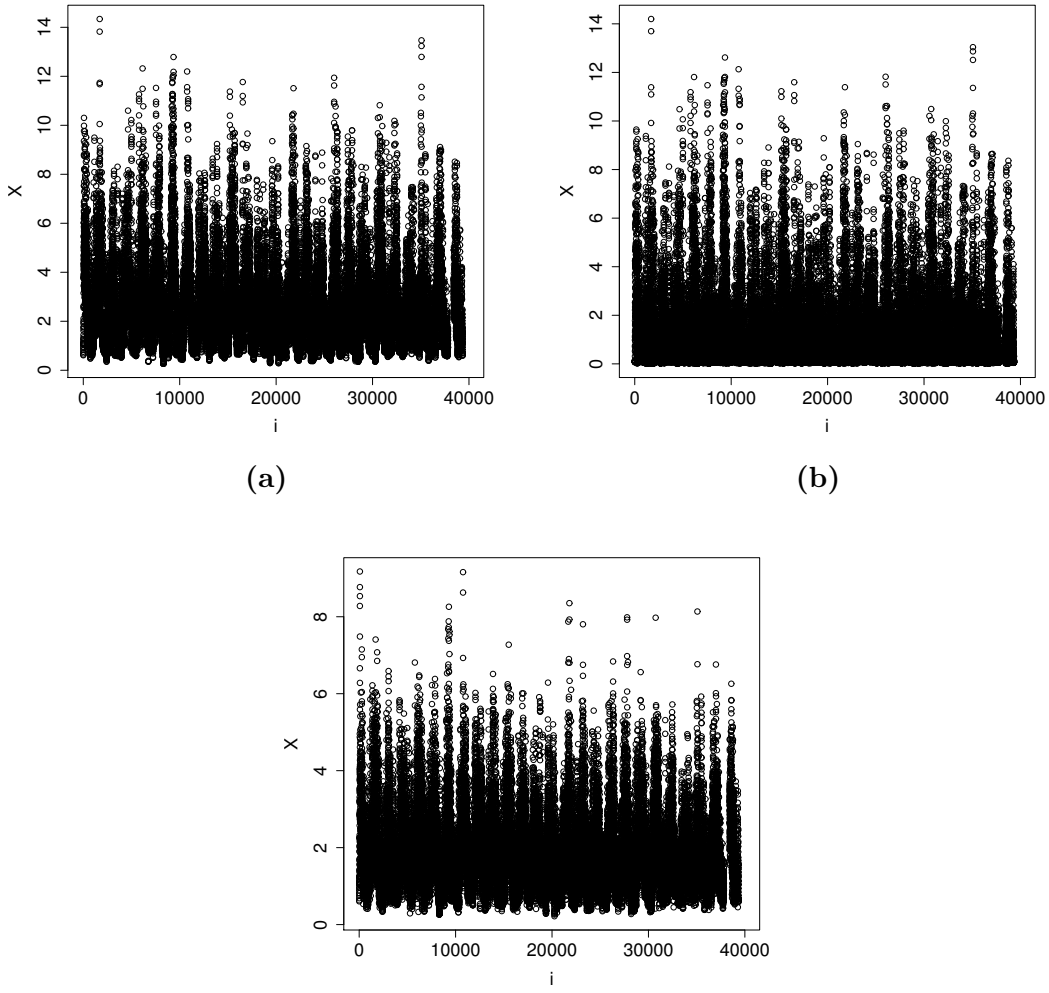
We have observations from two different locations on the Norwegian continental shelf. The first location is the oil field "Draugen", which is situated in the Norwegian Sea, about 145 kilometres to the north-west of Kristiansund. The other is the oil field "Ekofisk", which is situated in the middle of the North Sea, between Scotland and Denmark.

For each of the two oil fields, we have 12 years and 8 months of daily observations, from January 1, 1990, to August 11, 2002. The observations were made every three hours. Thus, there are eight observations for each day and 2920 for each year, except in the last year. There are 36,919 observations in all. All the numbers are given in metres $[m]$, with three decimals. Hence, these data are much more accurate than the wind speed data.

There will be three different vectors of wave height observations, X_{H_S}, $X_{H_{SWS}}$, and $X_{H_{SSW}}$, for the significant wave data, the wind sea data, and the swell data, respectively. We will have one set of vectors for the Draugen data, and another for the Ekofisk data. In figures 5 and 6 we see plots of the three vectors for the Draugen and the Ekofisk data, respectively. We note that the H_S plots and the H_{SWS} plots are very similar, especially for the high wave height values. The H_{SWS} height must, as mentioned, be smaller than the H_S height, but it seems to be only a little smaller than the H_s height for high wave

heights. We should therefore expect the return values of the H_{SWS} wave heights to be only a little smaller than the return levels of the H_S wave heights. The H_{SSW} values are seen to be much lower than the two others.

The same tendencies can be found in tables 3 and 4, where the yearly maxima of the Draugen and Ekofisk maxima are listed. Again, we see that the maximum H_{SSW} values are only slightly smaller than the corresponding H_S values. We also note that the Draugen numbers generally seem to be larger than the ones from Ekofisk.



(a) **(b)**

(c)

Figure 5: Plot of the Draugen **(a)** H_S data, **(b)** H_{SWS} data, and **(c)** H_{SSW} data.

A plot of only one year of observations, for example of the first year of the Draugen

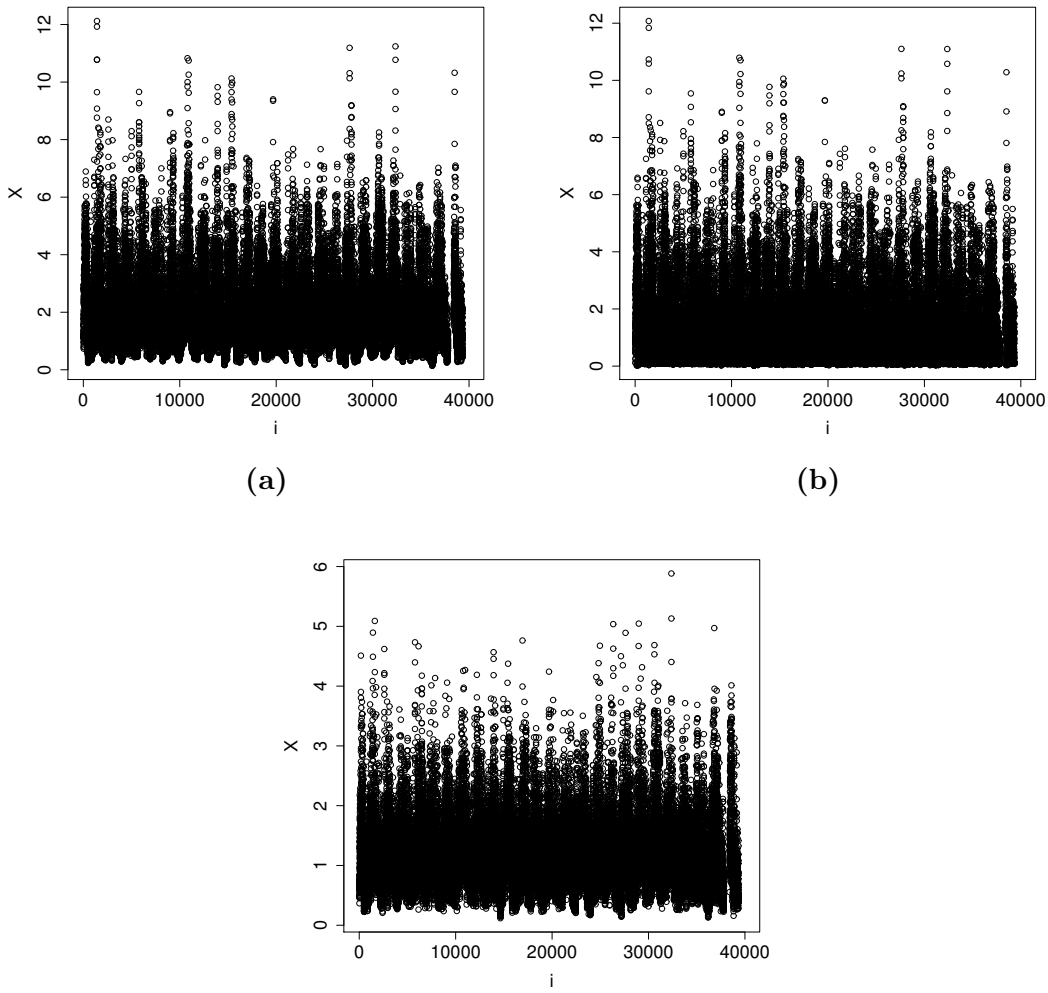


Figure 6: Plot of the Ekofisk **(a)** H_S data, **(b)** H_{SWS} data, and **(c)** H_{SSW} data.

H_S	year	1	2	3	4	5	6	7
	max. height	14.339	11.261	12.321	12.787	11.376	11.766	9.352
	year	8	9	10	11	12	13	
	max. height	11.513	11.941	10.298	10.819	13.468	9.113	
H_{SWS}	year	1	2	3	4	5	6	7
	max. height	14.199	11.187	11.809	12.615	11.225	11.595	9.289
	year	8	9	10	11	12	13	
	max. height	11.391	11.822	10.207	10.493	13.039	8.647	
H_{SSW}	year	1	2	3	4	5	6	7
	max. height	9.176	6.810	6.532	9.159	6.513	7.274	5.910
	year	8	9	10	11	12	13	
	max. height	8.353	6.838	7.985	7.975	8.136	6.759	

Table 3: The yearly maxima of the Draugen ocean wave data.

H_S	year	1	2	3	4	5	6	7
	max. height	12.122	9.659	8.955	10.830	9.822	10.125	9.400
	year	8	9	10	11	12	13	
	max. height	7.128	7.962	9.196	11.241	6.518	10.323	
H_{SWS}	year	1	2	3	4	5	6	7
	max. height	12.076	9.542	8.901	10.791	9.770	10.058	9.298
	year	8	9	10	11	12	13	
	max. height	6.664	7.918	9.096	11.096	6.436	10.283	
H_{SSW}	year	1	2	3	4	5	6	7
	max. height	5.090	4.733	4.666	4.270	5.568	4.763	4.242
	year	8	9	10	11	12	13	
	max. height	3.556	5.037	5.046	5.884	3.715	4.013	

Table 4: The yearly maxima of the Ekofisk ocean wave data.

H_S data, as given in figure 7, reveals the same periodic structure as we found in the wind speed data, which is very natural, since ocean waves are generated by wind. A plot of a smaller section of the data, for example the very first month of the Ekofisk H_S data, as given in figure 8, reveals a correlation among the observations. As for the wind speed data, we must take account for this correlation during our analysis.

Again, we can assume that the extreme value distribution is of the Gumbel form. By a similar argument as was used for the wind speed data, we can also assume that the constant b in (22) must be nonnegative, $b \geq 0$. [14]

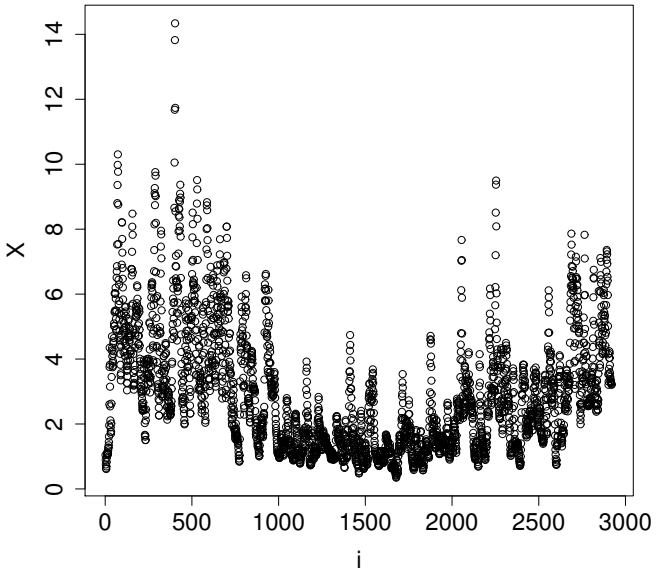


Figure 7: Plot of the first year of the Draugen H_S data.

4 Results

4.1 The Ørlandet Wind Speed Data

To illustrate the use of the AER method in practice, we are going to go rather thoroughly through the analysis of the Ørlandet wind speed data. In all our examples, we will use the 100 years return value estimate to compare our results.

Our wind speed data are originally given in a long vector X with 7305 elements, each element representing the FFX observation of one day. In the data set, there are 15 missing values, given as "NA". Since we wish to find the return levels for long time intervals, it is natural to use a year as our time period. Further, having 20 years of data,

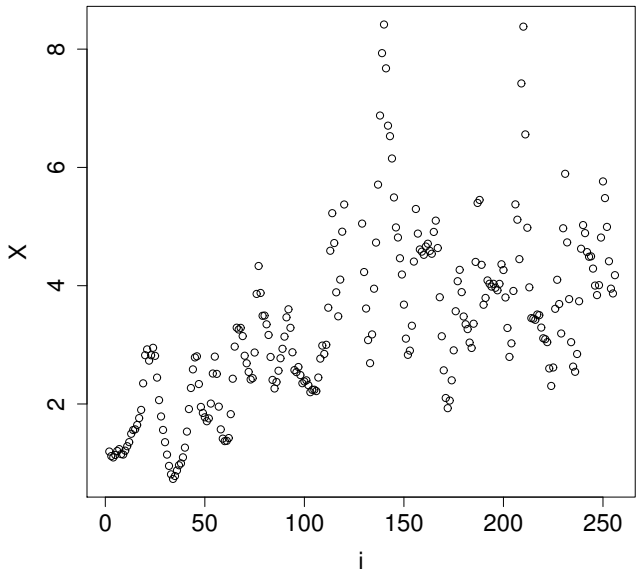


Figure 8: Plot of the first month of the Ekofisk H_S data.

it is natural to let each year represent a block of data. We will then have $R = 20$ blocks with $N = 365.25$ observations per time period.

It is convenient to work on a a non-dimensional scale. Therefore, we transform our scale η into the non-dimensional scale $\frac{\eta}{\hat{\sigma}}$, where $\hat{\sigma}$ is the empirical standard deviation of the entire data series in consideration. We will refer to this transformed scale simply as η, and in the future all plots and numbers will be given using this scale. The return levels, however, will be transformed back to the original scale. In this case, $\hat{\sigma} = 3.5931$.

First, we will have to discretize the η scale. As noted in section 3.1, the wind speed data were heavily discretized. Now, from (26) and (28) it follows that using a too fine discretization when finding estimates $\hat{\epsilon}_k(\eta)$ will only give us redundant information. But in section 3.1 we saw that the observation values were discretized with a maximum difference of 0.5. So, we should use that discretization to avoid too much redundant information. On the transformed scale, such a discretization corresponds to a step $d_\eta = 0.14$ between the η levels.

We now proceed to find the estimates $\hat{\epsilon}_k(\eta)$. We have not yet decided which value of k to use, but calculating the estimates $\hat{\epsilon}_k(\eta)$ for several values of k and comparing them will often give us the information we need to make the decision.

Finding the estimates is rather straightforward. We simply calculate the $\hat{\beta}_k(\eta)$ values for each of the R blocks and use the average of those numbers as our estimate, as given by (30). When finding the standard deviations, we use (31). For $k \geq 2$, when estimating

$\bar{\epsilon}_k(\eta)$, we can use either of the two formulae (29) or (28). A comparison of the $\log \hat{\epsilon}_2(\eta)$ estimates for using the two formulae is given in figure 9. As we can see, the estimates are almost identical for large values of η, while they diverge for smaller η. However, since the uppermost curve, the curve of estimates made from (28), is more correct for smaller values of η, we may be able to have a slightly longer fit if we use that curve. Therefore, in the following analysis, the formula (28) has been used.

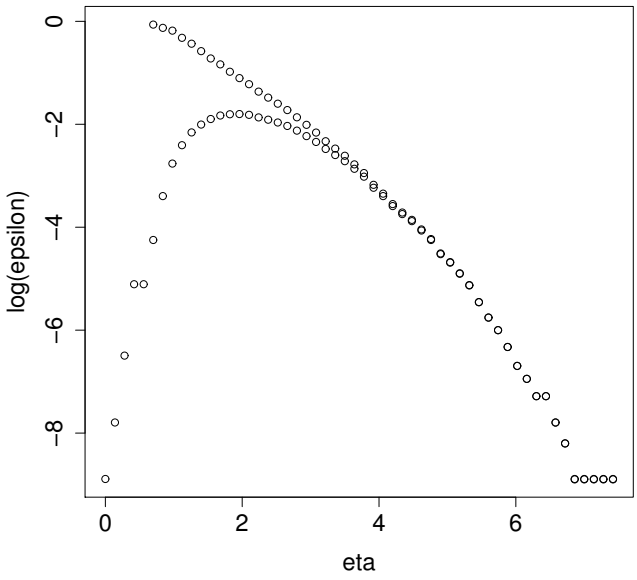


Figure 9: Plot of $\log \hat{\epsilon}_2(\eta)$ for the Ørlandet wind speed data, calculated by (28) (starting in the upper left corner) and by (29) (starting in the lower left corner).

The estimates $\log \hat{\epsilon}_k(\eta)$ for values of k from 1 to 6 calculated using (28) are shown in figure 10. As we can see, the estimates for $k \leq 2$ are nearly identical, at least for η values greater than about $\eta = 5$, and are not too far from each other for smaller values of η. On the other hand, the line representing $k = 1$ is seen to be at some distance from the others, at least for η values smaller than about $\eta = 5$, and again for η values greater than about $\eta = 6$. The fact that the estimates $\log \hat{\epsilon}_k(\eta)$ for $k \geq 2$ seem to converge, but are different from the estimates for $k = 1$, indicates that setting $k = 2$ will account for the correlation in the data series. Generally, k should be as small as possible, as choosing a larger value of k than is necessary will only give us more inaccurate results, because the estimates will be made with fewer conditional exceedances. Although it seems clear that $k = 2$ is the optimal value of k, we will here do the analysis for several values of k, to compare the results. We will start with $k = 1$, the case which will serve as our main example of the AER method.

Having chosen a k value, we should make a log plot of the estimates with the 95%

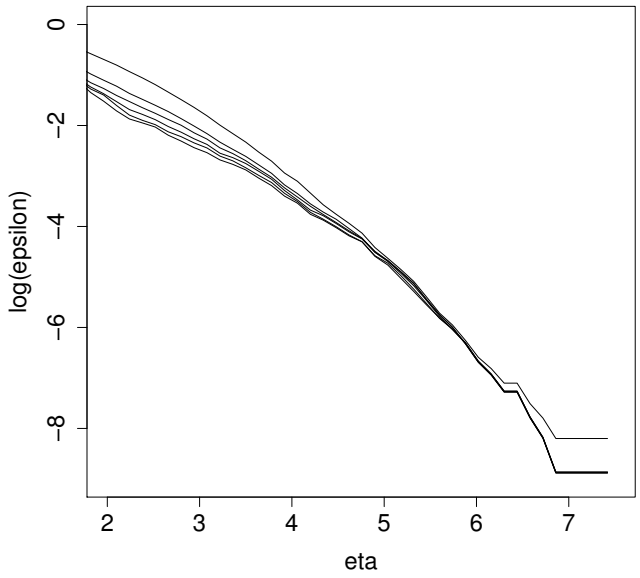


Figure 10: Convergence plot of $\log \hat{\epsilon}_k(\eta), k = 1, \ldots, 6$ for the Ørlandet wind speed data. The uppermost curve represents $k = 1$; then $k = 2$ follows beneath, then $k = 3$, *etc.*

confidence intervals $\hat{\mathrm{CI}}\big(\log \bar{\epsilon}_k(\eta)\big)$ added. Hopefully, from that plot we will be able to read the approximate position of the optimal values of η_1 and η_2. For $k = 1$, such a plot is shown in figure 11.

From this figure we must choose bounds within which we would expect η_1 and η_2 to be found. Looking at the plot, we notice that the confidence intervals widen significantly for $\eta > 6$. Hence, since we wish to avoid the more uncertain $\hat{\epsilon}_1(\eta)$ estimates, it seems right to cut away the estimates where η is larger than about 6. Using $(5, 7)$ as our $(\eta_2^{(low)}, \eta_2^{(high)})$ interval, we should be fairly sure to include the optimal value of η_2. When it comes to η_1, reading information from the plot is more difficult. The optimal value will probably be found where the curve constituted by the $\hat{\epsilon}_1(\eta)$ points straightens out, somewhere near $\eta = 1$. We should be on the safe side using $(0, 2)$ as our $(\eta_1^{(low)}, \eta_1^{(high)})$ interval.

As mentioned, our discretization of the η scale is quite coarse. Actually, it turns out that there are 15 η points in the selection $(0, 2)$, while there are 14 in the selection $(5, 7)$. These numbers are low enough that we can conveniently compare all the possible combinations of η_1 and η_2 values, without using only a representative selection, as proposed in section 2.6. Thus, we only need to do the analysis once. All in all, there will be $15 \cdot 14 = 210$ combinations of η_1 and η_2 values to be computed, and 50 different lengths $L = \eta_2 - \eta_1$ for which to compare the Δ values of the linear fits made using those combinations.

As mentioned in section 2.6, we must decide on which are possible values of the

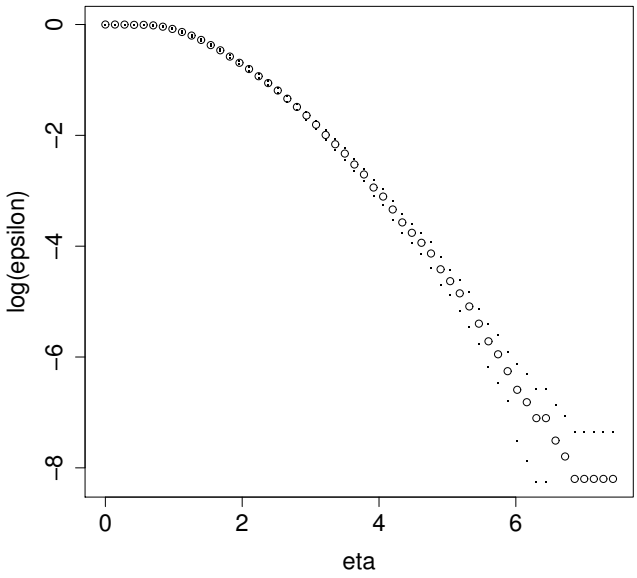


Figure 11: Plot of the estimates $\log \hat{\epsilon}_1(\eta)$ (circles) for the Ørlandet wind speed data, with 95% confidence interval bands (dots).

constants q and b. A lower bound for q is given by $\max_{\eta_1 \leq \eta \leq \eta_2} \hat{\epsilon}_k(\eta)$, and will differ with each combination of η_1 and η_2, as different subsets of the set of estimates $\hat{\epsilon}_1(\eta)$ are used. We must decide on an upper bound of q ourselves. A little trial and error seems to indicate that by setting $q^{(high)}$ to $q^{(high)} = 1.4$, all likely values of q should be covered. Generally, if our analysis return fits where $q = q^{(high)}$, we should raise its value and try again. On the other hand, if the good fits are found to have q values that are significantly smaller than $q^{(high)}$, we can lower it to get more precise results. The discretization of the q values, determined by d_q, must be decided on so as to obtain the desired accuracy. In this case, $d_q = 0.01$ is found to be sufficiently small.

The upper bound of the possible b values was found to be $\min_{\eta_1 \leq \eta \leq \eta_2} \eta = \eta_1$. In this case we also have a given lower bound $b^{(low)} = 0$, since, as was mentioned in section 3, we cannot have negative b values. As for the discretization of the b values, determined by d_b, it was said in section 2.6 that it does not need to be as small as d_q, since the value of q is much more important for the linearity of the fit. $d_b = 0.05$ is found to be sufficiently small.

We do the analysis of step 1 on a computer, and have the computer print the results in a table with the same number of rows as the number of differents lengths L. Hence, there will be 50 rows. In each row of the table it is convenient to print the length L and the corresponding η_1 and η_2 values, the corresponding Δ value of the best fit, as well as the corresponding constants q, b, a and c of that fit, and the corresponding estimate of

the 100 years return level x_{100}. We should also make a log log-log plot and a log plot for each fit. Finally, we should make plots of the return level against L, η_1, η_2, and perhaps Δ.

For $k = 1$, a selection from the table of results is shown in table 5. The corresponding plots cannot all be shown here, of course, but three pairs of the log log-log plots and log plots are given in figure 13. The return level plots are shown in figure 12.

When trying to select the best fit from the 50 alternatives, we can exclude many of them at once. The ones with the highest Δ values, such as lines 26, 32, 47, 49, and 50, obviously are poor fits. Further, all the fits in lines 1–30, with the exception of line 26, all have very similar Δ values. If two fits have the same Δ values, the one with the larger L value will be preferred. Thus, we can exclude lines 1–29 from consideration. This leaves us with lines 30–46, and among those line number 41 seems to stand out. Its Δ value is not much higher than in the preceeding lines, about 0.02, but from the next line on the Δ values rise to 0.1 and 0.2. So, choosing the values of line number 41 seems to give us a fit where the Δ value is as small as possible, when at the same time the L value is kept as large as possible.

Next, we look at the return level plots to see if they are consistent with our choice. In figure 12 we have plotted the estimates of x_{100} for each fit against the corresponding L, η_1, η_2, and Δ values. The largest estimates where left out, so as not to distort the plot.

When plotting $\hat{x}_{100}$ against L, we expect the estimates of x_{100} to be fairly stable for L values lower than the optimal one. From plot **(a)** in figure 12, $L = 6$ seems to be a good value. As mentioned in 2.6, the return level estimates should also be stable for η_1 values higher, and for η_2 values lower than the optimal ones. From plots **(b)** and **(c)** in the same figure, the return level estimates seem to stabilize for η_1 values larger than about 0.8, and for η_1 values larger than about 7. The last of the plots in figure 12, plot **(d)**, where the estimates of x_{100} are plotted against Δ, is not very helpful in this case.

Our analysis of the plots in figure 12 seems to strengthen the conclusion drawn from table 5. The fit described by the numbers in line 41 of that table does indeed have $L = 6.02$, $\eta_1 = 0.84$, and $\eta_2 = 6.86$, numbers that fit well with our analysis of the return level plots. In this case, it is really simple.

Finally, we turn to the pairs of log log-log plots and log plots. Three of those are shown in figure 13, where the plots of lines 20, 41 and 45 are given as examples of how to interpret such plots.

The first plot, **(a)**, corresponds to line 20 in the table. The linear fit is good, with a Δ value of just 0.01999, and it seems to give a good fit in the log plot, but it is too short, as it is obvious from the log plot that η_2 could have been moved further to the right. In the third plot, **(c)** which corresponds to line 52 in the table, we have a large L value, but the Δ value is high, and the linear fit is not good at all. Neither is the fit in the log plot. A comfortable middle ground is found in the second plot, **(c)**, which corresponds to line 41 in the table. The log log-log plot fit gives a straight line, and the log plot fit is very good. In fact, it is seen from the log plot that η_2 could not possibly have been moved further to the right. All in all, our reasoning seems to indicate that the numbers

	L	Δ	η_1	η_2	q	b	a	c	x_{100}
1	3.08	0.01282	1.96	5.04	1.281	0	0.2848	1.76	28.28
2	3.22	0.01305	1.82	5.04	1.241	0.05	0.2894	1.757	28.25
					...				
15	4.06	0.01285	1.68	5.74	1.168	0.25	0.3384	1.694	28.43
16	4.2	0.01298	1.54	5.74	1.14	0.3	0.3475	1.685	28.43
17	4.2	0.01273	1.68	5.88	1.298	0	0.2915	1.747	28.36
18	4.34	0.0129	1.4	5.74	1.117	0.35	0.3588	1.673	28.45
19	4.34	0.01264	1.54	5.88	1.28	0	0.2861	1.757	28.31
20	4.48	0.01278	1.4	5.88	1.257	0.05	0.297	1.742	28.36
21	4.62	0.01467	1.54	6.16	1.25	0	0.2742	1.782	28.13
22	4.62	0.01324	1.26	5.88	1.189	0.15	0.3073	1.736	28.27
23	4.76	0.0144	1.26	6.02	1.239	0	0.271	1.788	28.11
24	4.76	0.015	1.54	6.3	1.25	0	0.274	1.782	28.12
25	4.9	0.01473	1.26	6.16	1.239	0	0.2709	1.788	28.1
26	4.9	1.209	0.14	5.04	1.4	0	0.6957	0.9843	58.47
27	5.04	0.01702	1.12	6.16	1.074	0.4	0.3553	1.689	28.21
28	5.04	0.01504	1.26	6.3	1.239	0	0.2708	1.789	28.09
29	5.18	0.02199	0.98	6.16	0.9934	0.65	0.4348	1.605	28.44
30	5.18	0.01736	1.12	6.3	1.074	0.4	0.3552	1.69	28.2
31	5.32	0.02205	0.98	6.3	0.9934	0.65	0.4347	1.606	28.43
32	5.32	1.284	0.14	5.46	1.4	0	0.697	1.023	52.51
33	5.46	0.02287	1.4	6.86	1.037	0.55	0.4118	1.622	28.51
34	5.46	0.02274	0.84	6.3	0.9815	0.7	0.4583	1.58	28.57
35	5.6	0.02291	1.26	6.86	1.079	0.45	0.3863	1.645	28.48
36	5.6	0.023	1.12	6.72	1.014	0.6	0.4232	1.612	28.51
37	5.74	0.02296	1.12	6.86	1.034	0.55	0.4103	1.624	28.5
38	5.74	0.02405	0.84	6.58	0.9815	0.7	0.4583	1.579	28.6
39	5.88	0.02521	0.98	6.86	0.9934	0.65	0.4349	1.604	28.47
40	5.88	0.02419	0.84	6.72	0.9815	0.7	0.4583	1.579	28.6
41	6.02	0.02367	0.84	6.86	0.9815	0.7	0.4583	1.579	28.59
42	6.02	0.2212	0.56	6.58	1.064	0	0.1778	2.037	26.68
43	6.16	0.1055	0.7	6.86	1.105	0	0.2072	1.939	27.33
44	6.16	0.4694	0.42	6.58	1.024	0	0.1368	2.209	25.66
45	6.3	0.2152	0.56	6.86	1.064	0	0.1785	2.03	26.82
46	6.44	0.4687	0.42	6.86	1.024	0.05	0.1552	2.133	26.13
47	6.44	1.459	0.14	6.58	1.4	0	0.6936	1.11	42.78
48	6.58	0.7409	0.28	6.86	1.008	0	0.1148	2.318	25.22
49	6.58	1.478	0.14	6.72	1.4	0	0.6927	1.119	42
50	6.72	1.497	0.14	6.86	1.4	0	0.6917	1.127	41.24

Table 5: A selection of the results of the initial analysis of the Ørlandet wind speed data, with $k = 1$.

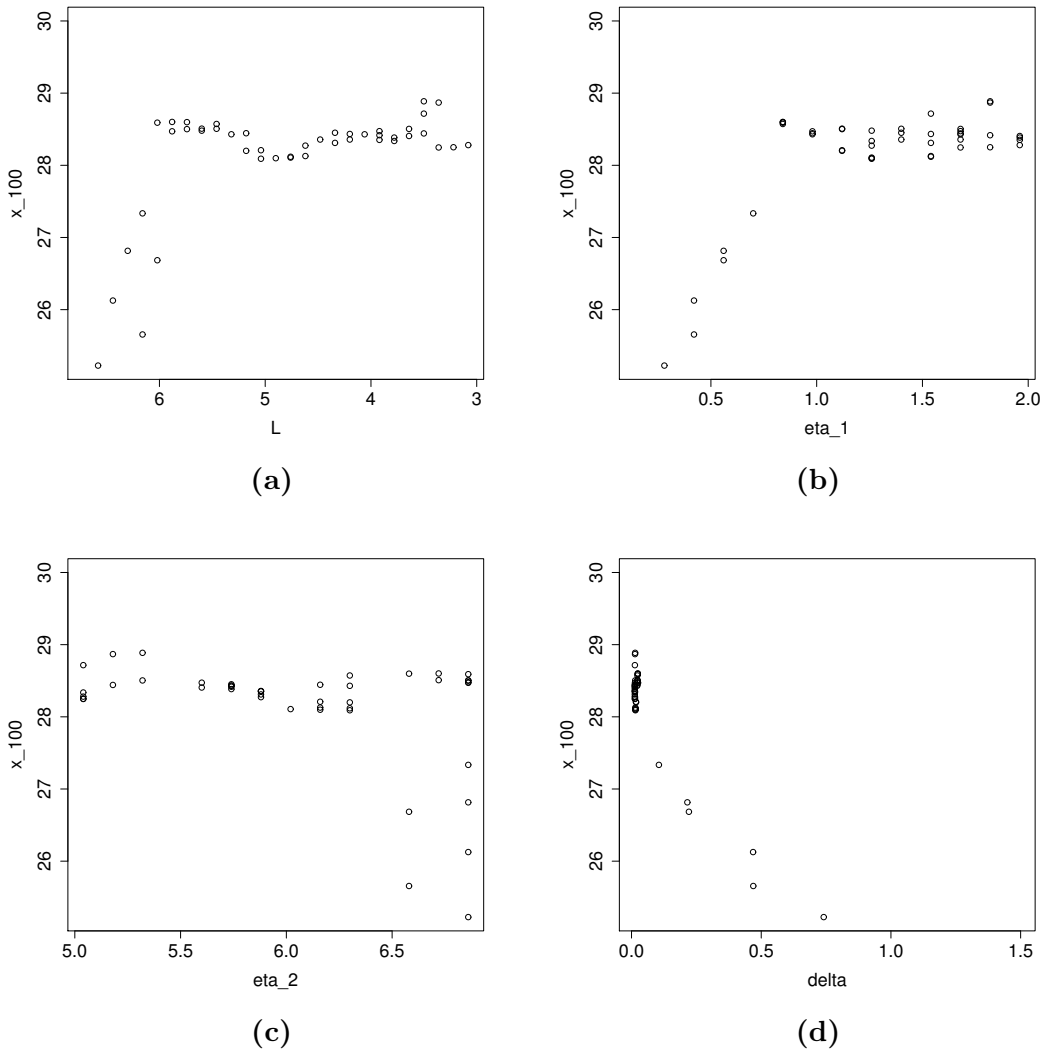


Figure 12: Step 1 performed on the Ørlandet wind speed data with $k = 1$. **(a)** Plot of $\hat{x}_{100}$ against L. **(b)** Plot of $\hat{x}_{100}$ against η_1. **(c)** Plot of $\hat{x}_{100}$ against η_2. **(d)** Plot of $\hat{x}_{100}$ against Δ. Some return level estimates have been left out.

given by row number 41 in table 5 are the most suitable for further analysis.

Of course, this was a very simple example. In more complex cases, analysing the table can be difficult, and the return level plots may not give such explicit information as it did here. Sometimes they may not be useful at all. However, looking at the fits in the plots is always useful, as it is instantly made clear if the numbers given in the table represent a good fit or not. Thus, the fitted curve in the log plot should always be the last point of reference. In the end, if we have a reasonably good fit in the log plot, it can be used for further analysis.

Choosing fit number 41 from the table gives us $\eta_1 = 0.84$ and $\eta_2 = 6.58$, while we will have initial estimates $\hat{q}_i = 0.9815$, $\hat{b}_i = 0.700$, $\hat{a}_i = 0.4583$, and $\hat{c}_i = 1.5792$. We now proceed to step 2 of the analysis, the final estimation of the constants, using the initial estimates as our starting values. We are going to fit the curve given by (22) to the estimates $\log \hat{\epsilon}_1(\eta)$, for $\eta_1 \leq \eta \leq \eta_2$, as described in section 2.8. For this task, we prefer using the Marquardt-Levenberg algorithm, through the *lsqnonlin* method in Matlab.

We need to have upper and lower bounds for the constants q, b, a and c. In fact, we have already given bounds for q and b. However, as was mentioned in section 2.8, only one of those variables needs to be bounded on both sides. Since b is, we can let q vary freely upwards. a and c only need to be positive. For q, b, a, and c we therefore set the lower bounds to be $\max_{\eta_1 \leq \eta \leq \eta_2} \hat{\epsilon}_1(\eta)$, $b^{(low)} = 0$, 0 and 0, respectively, and the upper bounds to be ∞, η_1, ∞ and ∞, respectively. As weights we use the $w_1(\eta)$ formula given by (43).

Now, running the method in Matlab, we obtain the results given in table 6. The initial estimates are given in the first row of the table, while the final estimates are given in the second row. We also give the corresponding errors of the fits, which are the values of (42).

	η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	error
init.	0.84	6.86	0.9815	0.700	0.4583	1.5792	28.5908	$1.771119 \cdot 10^{-3}$
final	0.84	6.86	0.9984	0.6262	0.4264	1.6114	28.4828	$1.584943 \cdot 10^{-3}$

Table 6: Final results for the Ørlandet wind speed data, with $k = 1$, compared to the initial estimates.

As we can see from the errors in the last column of the table, our new fit is slightly better than the one defined by the initial estimates. We also notice that the value of $\hat{q}$ changed only slighly, to become closer to 1. On the other hand, the b value has changed rather significantly, which we should perhaps have expected, since we used a rather coarse discretization of the b values, with $d_b = 0.05$. The $\hat{a}$ and $\hat{c}$ values have been slightly altered, and the 100 years return level has been adjusted slightly downward. However, there were no drastic changes, as step 2 is only a refinement of the results obtained in step 1.

Having found the final estimates of the constants, we proceed to find a 95% confidence interval for x_{100}, following the procedure sketched in section 2.10. First, the confidence bands of $\log \bar{\epsilon}_1(\eta)$ are transposed onto our new fit $\hat{f}(\eta)$, as shown in figure 14. Now,

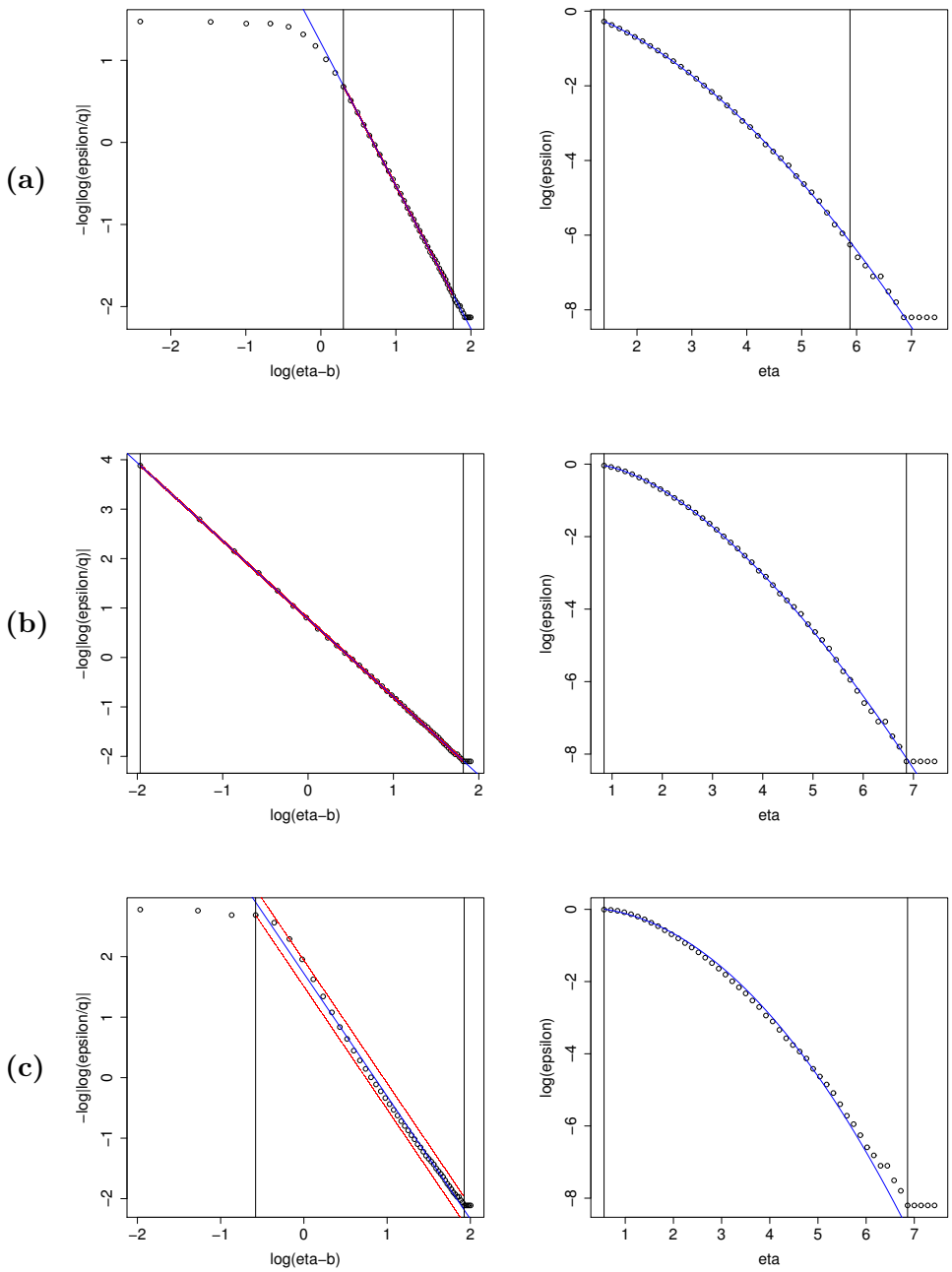


Figure 13: Three examples of plots made from table 5. To the left, log log-log plots. To the right, log-plotts. $\eta = \eta_1$ and $\eta = \eta_2$ are shown as vertical lines. The examples correspond to lines **(a)** 20, **(b)** 41, and **(c)** 45 of the table. The fitted curves are drawn in blue. The dots in red are at a distance Δ from the linear fits.

comparing with figure 11, we clearly see that the confidence bands are smoother, and thus more fitting for our purpose.

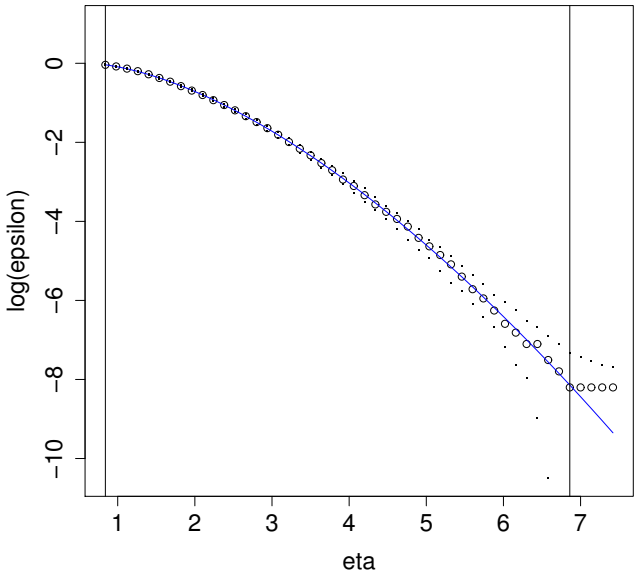


Figure 14: The 95% confidence bands of $\log \bar{\epsilon}_1(\eta)$ (dots) put onto the fit $\hat{f}(\eta)$ (the blue curve), for the Ørlandet wind speed data. Compare with figure 11. The left and right vertical lines represent η_1 and η_2, respectively.

Before calculating the confidence intervals, we should, as mentioned in section 2.10, make sure that there are no irregularities in the transposed confidence bounds which would distort the estimated confidence intervals for the return levels. In this case, we find that the confidence intervals on the logarithmic scale cannot be calculated for the two largest η values, since in those cases $\hat{q}\exp\{-\hat{a}(\eta-\hat{b})^{\hat{c}}\} - \frac{1.96\hat{s}_k(\eta)}{\sqrt{R}}$ will be negative, so that (45) cannot be used. We solve this problem by simply removing the confidence intervals for those two η values from our analysis, that is, for $\eta > 6.58$.

To find the confidence intervals, we again use Matlab, this time with the method *fmincon*. As for the final estimation of the constants, we need to put bounds on the variation of those constants. We can use the same bounds as above. Matlab gives us the results shown in table 7.

A plot showing how the curves corresponding to the confidence intervals fit in between the confidence bands of $f(\eta)$ is given in figure 15. The horizontal line represents the $\log \bar{\epsilon}_1(\eta)$ value corresponding to the 100 years level, and the 100 years level is found where the fitted curve from figure 14 crosses this line. The confidence bounds of x_{100} on both sides of $\hat{x}_{100}$ are found where the two outermost curves cross the same line. We see how both the upper and the lower confidence bounds have a corresponding curve (given

	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$
upper bound	1.1045	0	0.1989	1.9902	26.4867
estimate	0.9984	0.6262	0.4264	1.6114	28.4828
lower bound	0.9624	0.8400	0.5651	1.4417	30.2225

Table 7: 95% confidence interval for the 100 years return value of the Ørlandet wind speed data, with $k = 1$, with the corresponding constants.

in red in the plot) which neatly fits in between the 95% confidence bands on both sides of $\log \hat{\epsilon}_1(\eta)$. These curves are represented by corresponding values of q, b, a, c, which are called $q^{(high)}$, $b^{(high)}$, $a^{(high)}$, $c^{(high)}$, and $q^{(low)}$, $b^{(low)}$, $a^{(low)}$, $c^{(low)}$, respectively. Those are also given in table 7. The 95% confidence interval is $(26.49, 30.22)$.

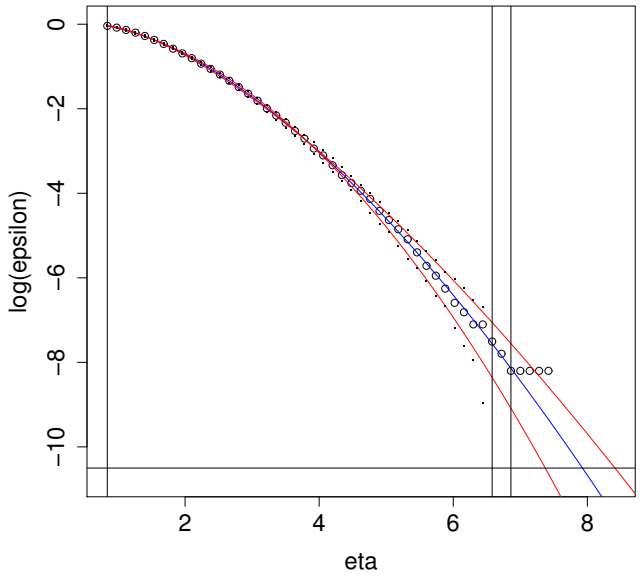


Figure 15: The 95% confidence intervals of the 100 years return level of the Ørlandet wind speed data. The $\log \bar{\epsilon}_1(\eta)$ value corresponding to the 100 years level is shown as a horizontal line. The leftmost and the rightmost vertical line shows η_1 and η_2, respectively. The middle horizontal line shows where the confidence bounds of the curve $f(\eta)$ have been cut. The fit $\hat{f}(\eta)$ is shown in blue, while the curves corresponding to the confidence interval limits are shown in red.

It may be of interest to estimate the return levels and confidence intervals using the squared weights formula $w_2(\eta)$ of (44) instead of $w_1(\eta)$ of (43). We then get the results shown in table 4.1. As we can see, the return level is almost the same as with the $w_1(\eta)$ weights formula, but slightly higher. This may be explained by the fact that the $w_2(\eta)$

η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	95% $\hat{\text{CI}}(x_{100})$
0.84	6.86	0.9832	0.6960	0.4595	1.5735	28.72165	(26.6783, 30.3781)

Table 8: Result of the analysis for the Ørlandet wind speed data, $k = 1$, with squared weights. Compare with tables 6 and 7.

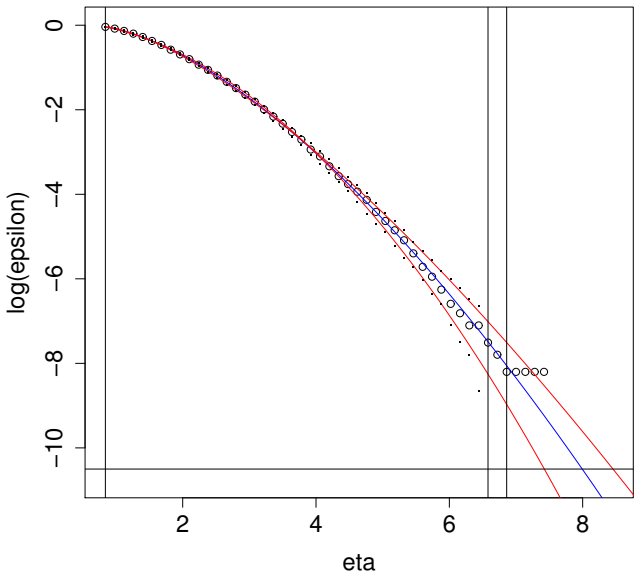


Figure 16: Result of the analysis for the Ørlandet wind speed data, $k = 1$, with squared weights. Compare figure 15.

formula does not put as much emphasis on the tail as $w_1(\eta)$ does. In this case it did not matter much, but in other cases this may lead to larger differences in the results. The fit is shown in figure 16.

We now go on to estimate return levels for $k \geq 2$. The analysis will be fairly similar to the one we have just gone through for $k = 1$. We may use some of the information found in the $k = 1$ analysis to ease our work when analysing $k \geq 2$, as the η_1 values of $k \geq 2$ will probably be greater than the η_1 value of $k = 1$. This information saves us from running through too many possible η_1 values. The results are given in table 9 and visualized in figure 17. We notice how the value of $\hat{q}$ decreases with increasing k values, and how the η_1 value is much larger and the 95% confidence interval wider for $k \geq 2$ than for $k = 1$. We also notice the similarity in the results for $k \geq 2$. This is as expected, since the convergence plot of figure 10 showed us that the $\log \bar{\epsilon}_k(\eta)$ estimates were almost identical for large values of η. We therefore accept the $k = 2$ estimates as our final estimates. Our estimated 100 years return value will then be $\hat{x}_{100} = 28.18$,

which seems to be a reasonable value, if compared to the yearly observed maxima of table 1. Especially, this estimate seems to indicate that the two extraordinary values of the table are indeed outliers.

k	η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	95% $\hat{\text{CI}}(x_{100})$
1	0.84	6.86	0.9984	0.6262	0.4264	1.6114	28.4828	(26.4867,30.2225)
2	2.24	6.58	0.6191	0	0.1918	1.9209	28.1763	(25.7821, 31.6467)
3	2.24	6.72	0.4566	0	0.1521	2.0205	28.1168	(25.6706,31.6344)
4	2.24	6.72	0.3522	0	0.1202	2.1275	27.9625	(25.5617,31.6070)

Table 9: Results for the Ørlandet wind speed data.

Finally, we will try to use the alternative estimate formula $\tilde{\beta}_k(\eta)$ from (29) for $\hat{\epsilon}_k(\eta)$, instead of $\hat{\beta}_k(\eta)$ from (28). Setting $k = 2$, we will have estimates of $\log \bar{\epsilon}_2(\eta)$ as given in figure 9, and will have the results given in table 10 and in figure 18. We here notice the broader confidence bands on both sides of $\log \hat{\epsilon}_k(\eta)$, which result in a broader confidence interval for the 100 years return level x_{100}. The value of η_1 is found to be slightly higher, and because the estimates of $\log \hat{\epsilon}_2(\eta)$ give a more rounded fitted curve, the return level estimate is lower, at 27.33, compared to 28.18 when using (28). The latter seems to be the better result, considering the lower value of η_1 and the width of the confidence interval. However, figure 9 seems to indicate that we would get the same results from both formulae if we chose a slightly larger η_1 value. Therefore, letting η_1 be as small as possible is probably not wise when using the $\tilde{\beta}_k(\eta)$ formula. Instead, one should choose the η_1 value after consulting the log plot with the added confidence bands, so as to avoid the uncertain $\bar{\epsilon}_k(\eta)$ estimates for smaller values of η. In this case, it seems that it would have been better to set η_1 somewhere near 3.5. Still, the $\tilde{\beta}_k(\eta)$ formula has the advantage that it involves less computation than the $\hat{\epsilon}_k(\eta)$ formula.

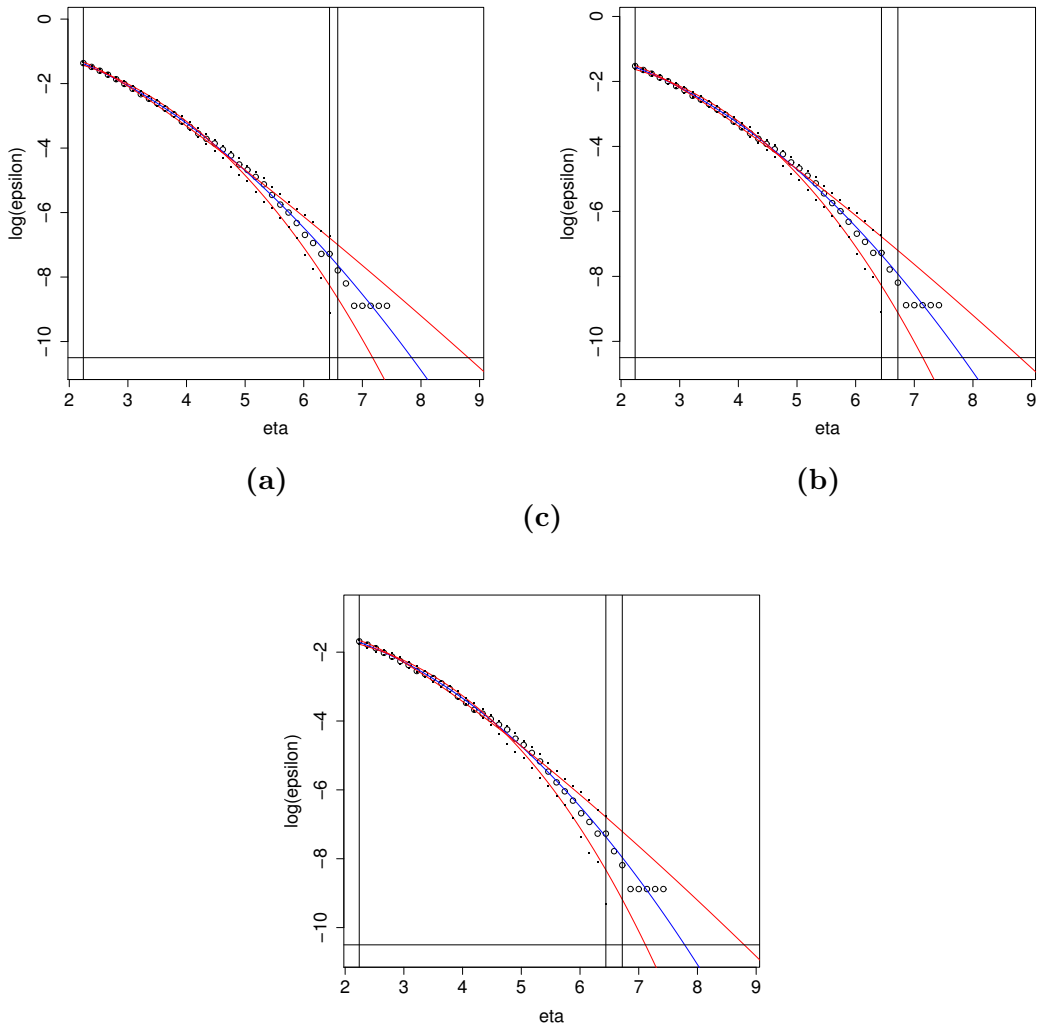
4.2 The Alta Wind Speed Data

We follow the same procedure for the Alta data as for the Ørlandet data. First, we make a convergence plot of $\log \hat{\epsilon}_k(\eta)$ with k values from 1 to 6, as shown in figure 19. The plot is similar to that of the Ørlandet data, but the difference between the line representing $\log \hat{\epsilon}_1(\eta)$ and the other lines is less marked. However, $k = 2$ seems to be a good choice for the Alta data as well.

When performing step 1 of the analysis, it turns out that the constant c is always very close to 1. The case $c = 1$ corresponds to the Gumbel distribution and will give

η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	95% $\hat{\text{CI}}(x_{100})$
2.38	6.72	0.2652	0.1159	0.0794	2.3586	27.3270	(24.7183,31.3453)

Table 10: Results for the analysis of the Ørlandet wind speed data with $k = 2$, using (29) instead of (28) in estimating $\bar{\epsilon}_2(\eta)$. Compare the second row of table 9.



(a) **(b)**

(c)

Figure 17: Results for the Ørlandet wind speed data. **(a)** $k = 2$, **(b)** $k = 3$, **(c)** $k = 4$.

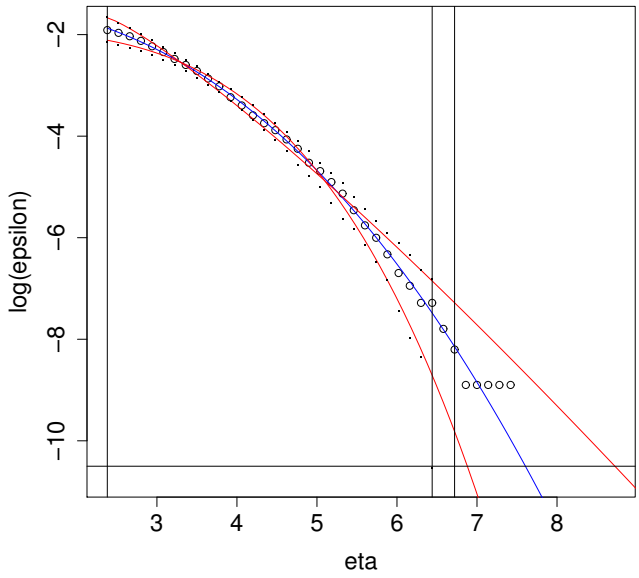


Figure 18: Results for the analysis of the Ørlandet wind speed data with $k = 2$, using (29) instead of (28) in estimating $\bar{\epsilon}_2(\eta)$. Compare plot **(a)** of figure 17.

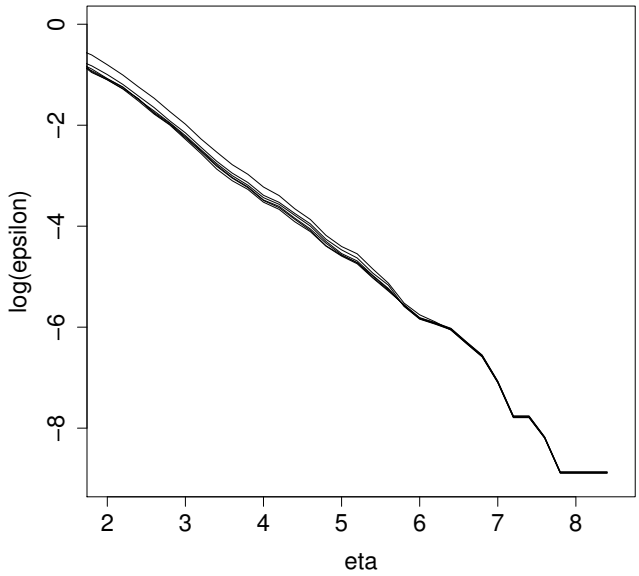


Figure 19: Convergence plot of $\log \hat{\epsilon}_k(\eta)$ for the Alta wind speed data. The uppermost line represents $k = 1$; then $k = 2$ follows beneath, then $k = 3$, *etc.*

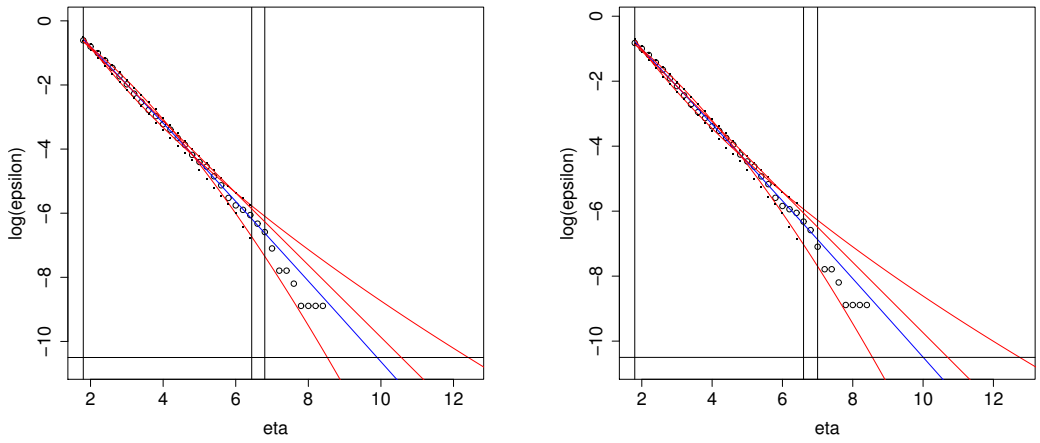
k	η_1	η_2		$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$
1	1.80	6.80	lower bound	1.4803	0	0.4248	1.5140	22.0966
			estimate	3.9661	0.0407	1.0791	1.0483	25.6527
			upper bound, $c \geq 1$	3.9760	0	0.0027	1.0000	27.3802
			upper bound	58.9684	0	3.2235	0.5998	32.0909
2	1.80	7.00	lower bound	1.1456	0	0.4065	1.5204	22.1886
			estimate	3.1979	0.0098	1.0618	1.0410	25.9359
			upper bound, $c \geq 1$	2.9816	0	1.0834	1.0000	27.7740
			upper bound	53.9521	0	3.3555	0.5747	33.0482
3	1.80	7.00	lower bound	1.0928	0	0.4164	1.5057	22.2355
			estimate	3.0931	0.0168	1.0935	1.0240	26.1263
			upper bound, $c \geq 1$	2.7539	0	1.0740	1.0000	27.7789
			upper bound	57.5959	0	3.4809	0.5587	33.5458

Table 11: Results for the Alta wind speed data.

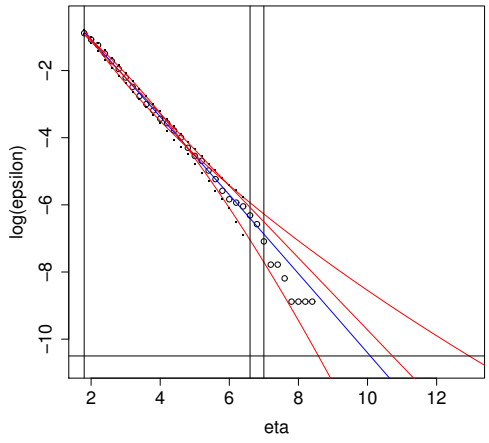
us a straight line in the log plot. Further, the fit can be described by infinitely many combinations of q and b. We therefore try to fix both q and c before continuing with step 2 of the analysis, setting $q = 1$ and $c = 1$ and varying only b and a. We then compare with the usual method, the one described in section 2.8 and used with the Ørlandet data. In fact, it turns out that the latter gives smaller errors for the fitted curve $\hat{f}(\eta)$. We will therefore use it for the Alta data as well.

During step 1, we had to set $q^{(high)}$ much higher than for the Ørlandet data and use a much larger d_q value. Indeed, the values of the constants are very different from that case. Table 11 shows the results after performing step 2. The fits are shown in figure 20. For each value of k, there are four rows in the table. The second row contains the estimated 100 years return value and the constants of the corresponding fit. The first and the fourth row contain the lower and upper confidence bounds of x_{100}, respectively. These are represented by the outermost red curves in the plots. The confidence intervals are very broad, a fact which is due to the low value of the constant c of the curve that defines the upper confidence bound. For example, for $k = 3$, we have $c = 0.5587$, which makes the curve turn outwards. But when we compare various Gumbel fittings of wind speed extremes, it is found that the c value stays above 1. Therefore, it may be justifiable to set a lower bound for the c values at $c = 1$ when finding the confidence intervals. Doing this, we obtain the upper confidence bounds of the third row, and much shorter confidence intervals. In fact, the new confidence intervals have half the length of the old. The new upper bound is represented by the middle red curve in the plots. This case illustrates the importance of considering the physical properties of the phenomenon under study.

The estimated 100 years levels are lower than the corresponding estimates for the Ørlandet data, which is reasonable, as the Alta yearly maxima were seen to be lower than the corresponding Ørlandet maxima. Although both the Alta and the Ørlandet data are wind speed observations from Norway, the curves in figure 20 and the ones in



(a) **(b)**



(c)

Figure 20: Results for the Alta wind speed data. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$. The middle red curve represents the curve giving the upper confidence bound when $c \geq 1$.

figure 17 are very dissimilar, and seems to indicate that the wind at Ørlandet behaves differently than the wind at Alta.

4.3 The Ocean Wave Data

The analysis of the ocean wave data is quite similar to the analysis of the wind speed data, and we will here only notice some points requiring special attention.

As noted in section 3, we have eight observations for each day throughout the year. As we are going to study long time intervals, it is again natural to use the year as our time unit. We therefore set $N = 365.25 \cdot 8 = 2922$ as the number of observations per period. We could of course use one block for each year, but that will give us only 13 blocks, one of which would be shorter than the others. $R = 13$ is somewhat too small to compute the confidence intervals. Instead, we choose to use two blocks per year, getting a total of $R = 26$ blocks. However, we cannot simply split the year in half at the middle, since the two blocks will be incomparable because of the periodicity of the data. Instead, we split each month of each year into two parts, assigning the first to the first block, and the second to the second block. Each block will then consist of 12 half-months, and will be a representation of the entire year. Now, we must remember that the length of the blocks is different from the number of observations per year, and use the appropriate formulae, as discussed in section 2.5. Also, we must remember that two of the blocks will be much shorter than the others, since we have only eight months of data from the last year. We therefore use a weighted mean and standard deviation in estimating $\hat{\epsilon}_k(\eta)$.

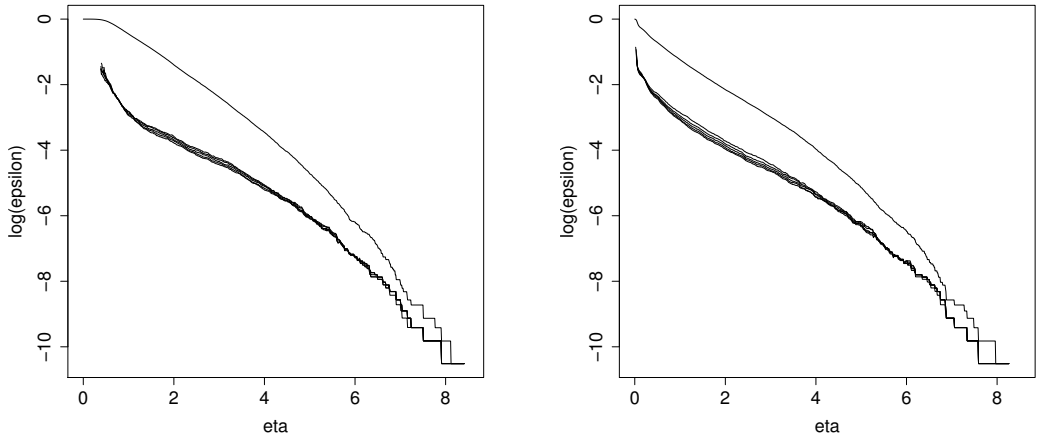
As for the wind speed data, in order to work on a non-dimensional scale, we transform η to the non-dimensional scale $\frac{\eta}{\hat{\sigma}}$, where $\hat{\sigma}$ is the standard deviation of the entire data series in consideration. Again, we refer to this transformed scale simply as η, but transform the return level estimates back to the original scale.

We choose to discretize the η scale with $d_\eta = 0.01$, although we could possibly have used $d_\eta = 0.001$. However, $d_\eta = 0.01$ gives a smaller and more convenient number of points to work with. In figures 21 and 22 we show convergence plots of the Draugen and Ekofisk data, respectively. In all these plots, the uppermost line represents $k = 1$, while the other lines, $k = 2, \ldots, 6$ converge to a single line. This indicates that setting $k = 2$ will again take care of the correlation in the data.

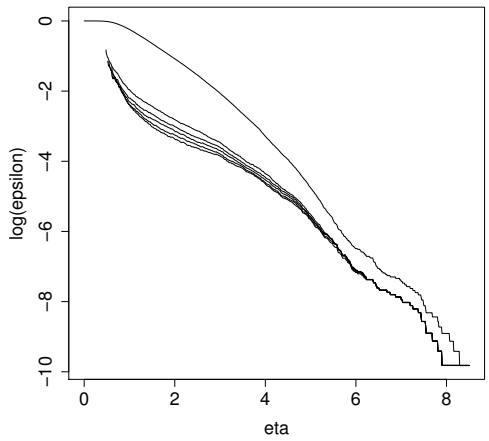
When inspecting the convergence plots in figures 21 and 22, we notice that the $\log \hat{\epsilon}_k(\eta)$ curves seem to have a bend for larger η values. This is especially obvious in the case of the H_{SSW} data from both the Draugen and the Ekofisk oil fields, but a smaller bend seems to occur in the other plots as well. This does not seem to be due to uncertainty among the estimates alone, but could indicate that the distribution of the ocean wave heights is inhomogeneous. In that case the AER method is invalid. Anyways, such a bend will make it more difficult to fit a curve to the $\log \bar{\epsilon}_k(\eta)$ estimates, and we will probably find that the η_2 values are small, as the bend must be left out.

For the wind speed data, we had such a small number of η values that it was enough to run step 1 of the method once. However, this time the number of possible η_1 and η_2 values will be much greater, since the discretization of η is finer, with $d_\eta = 0.01$. We therefore run step 1 twice, as described in section 2.6, and zoom in on the optimal η_1

and η_2 values. As was seen in section 3, b must be nonnegative, and therefore we set $b^{(low)} = 0$. To begin with, we set $q^{(high)}$ at 1.5, and use $dq = 0.01$. But in the cases $k \geq 2$, after the first run of step 1, we discover that the good fits exclusively have q values between 0.01 and 0.1. Therefore, before the second run, we change our $q^{(high)}$ value to 0.15 and the d_q value to 0.001, so as to have a more accurate q value.



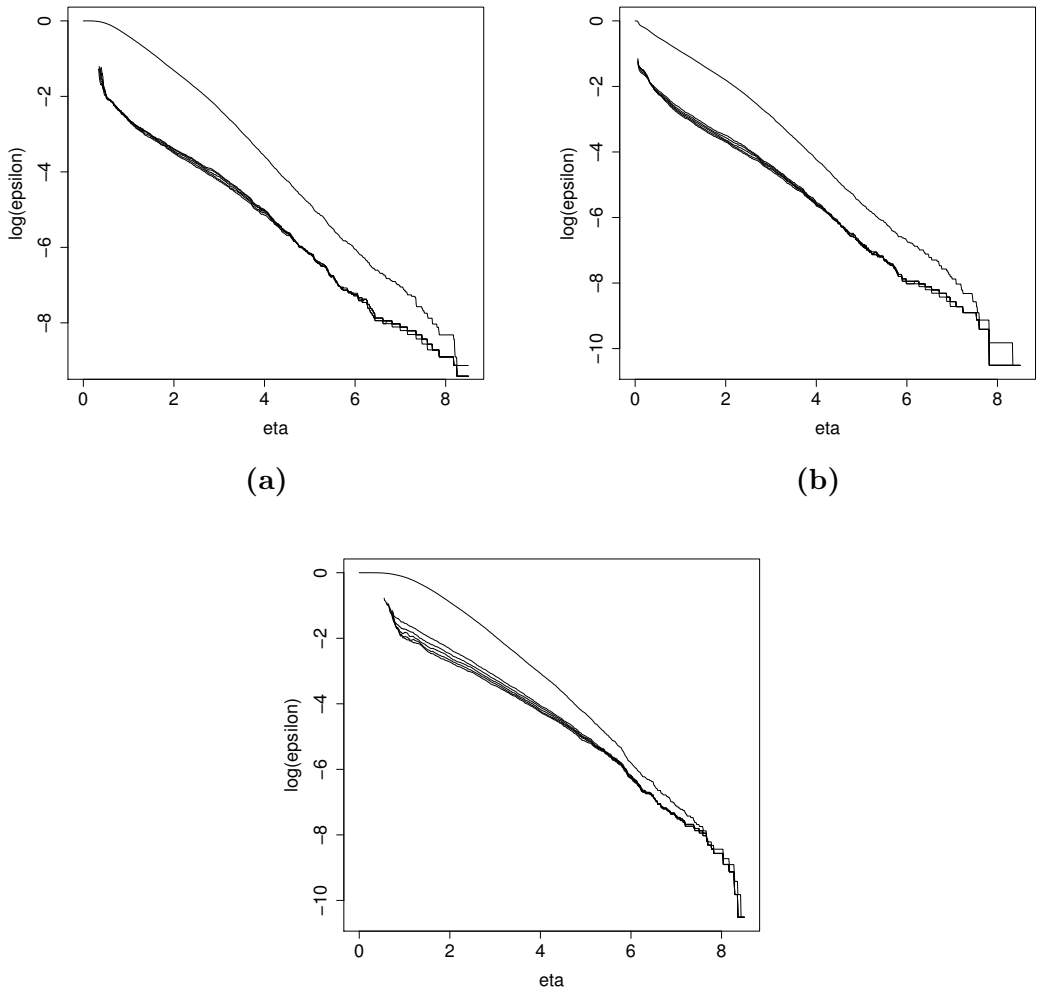
(a) **(b)**



(c)

Figure 21: Convergence plots of $\log \hat{\epsilon}_k(\eta)$ for the Draugen data, for $k = 1, \ldots, 6$, with $k = 1$ uppermost. **(a)** H_S, **(b)** H_{SWS}, **(c)** H_{SSW}.

We first look at the Draugen data. In table 12 we give the results of the analysis. We notice that for $k = 2$ and $k = 3$, we actually get higher return level estimates for



(a) **(b)**

(c)

Figure 22: Convergence plots of $\log \hat{\epsilon}_k(\eta)$ for the Ekofisk data, for $k = 1, \ldots, 6$, with $k = 1$ uppermost. **(a)** H_S, **(b)** H_{SWS}, **(c)** H_{SSW}.

the H_{SWS} data than for the H_S data. For example, for $k = 2$, we have that the H_S 100 years return level estimate is 15.8162, while the H_{SWS} return level estimate is 16.5806. Now, since the H_{SWS} data are observations of the wind sea, which constitutes a part of the significant wave height H_S, that should be impossible. The paradox may perhaps be explained by the fact that, as we saw from tables 3 and 4, the maximum values of the H_{SWS} data seem to be only slightly smaller than the maximum values of the H_S data, and that we should expect the return levels of the two wave classes to be close to each other. Thus, when taking the uncertainty of the estimates into account, a result such as ours may not be implausible. Indeed, the difference between the H_S and the H_{SWS} estimates are not very great when compared to the width of the corresponding 95% confidence intervals. On the other hand, we have seen from the log plots that the curves of $\bar{\epsilon}_k(\eta)$ seem to behave strangely for large η. Hence, it is probably wise to regard the results with a bit of scepticism.

When estimating the confidence intervals, we should plot the confidence bands on the curve $\hat{f}(\eta)$ beforehand, to see if some of them must be cut away. It turns out that in many cases it must. Figure 23 gives an example. In that case, the Draugen H_S data with $k = 2$, the confidence bands are found to be unreliable for η values larger than about 6.5. Using the bands for all $\eta_1 \leq \eta \leq \eta_2$ clearly will give us too short confidence intervals. Therefore, we should cut away the part $\eta > 6.5$ when doing the confidence interval analysis. As can be seen in table 12, the estimated confidence intervals are very broad. As for the Alta wind speed data, this is due to the values of c. By studying the properties of the ocean waves in general, we may be able to bound the c values and shorten the confidence intervals, as was done with the Alta data.

k	η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	$\hat{\text{CI}}(x_{100})$
				H_S				
1	2.00	6.71	0.6682	0	0.3538	1.5569	16.5390	(13.6420,21.2506)
2	2.29	7.21	0.0384	0	0.1144	1.9750	15.8162	(13.2189,20.8974)
3	2.29	7.21	0.0360	0.0196	0.1120	1.9820	15.8397	(13.2215,20.9512)
				H_{SWS}				
1	2.00	6.43	0.2510	0.0395	0.2572	1.6777	16.3646	(13.5801,21.3275)
2	1.90	6.73	0.0487	0.0280	0.2363	1.6346	16.5806	(13.4252,23.8859)
3	2.40	6.80	0.0347	0	0.1584	1.8123	16.1931	(12.8751,27.9530)
				H_{SSW}				
1	1.30	6.16	0.9938	0	0.3328	1.6550	9.5002	(8.6256,10.9632)
2	2.07	6.30	0.0890	0.0019	0.1023	2.1258	9.2083	(8.1307,11.3884)
3	2.29	6.30	0.0658	0.0107	0.0776	2.2467	9.1578	(8.0369, 11.6290)

Table 12: Results for the Draugen ocean wave data.

Plots of the fitted curves for the Draugen data are given in figures 24, 25, and 26. Here, we can clearly see what effect the bends of the $\log \bar{\epsilon}_k(\eta)$ estimate curves have had on the results. The H_S results seem to be good. For $k = 2$, a 100 years return level of 15.82 is not implausible, when we remember that the largest significant wave height

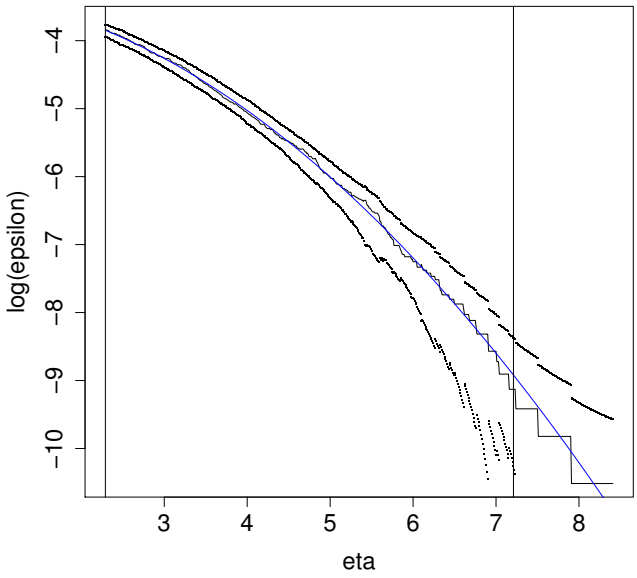
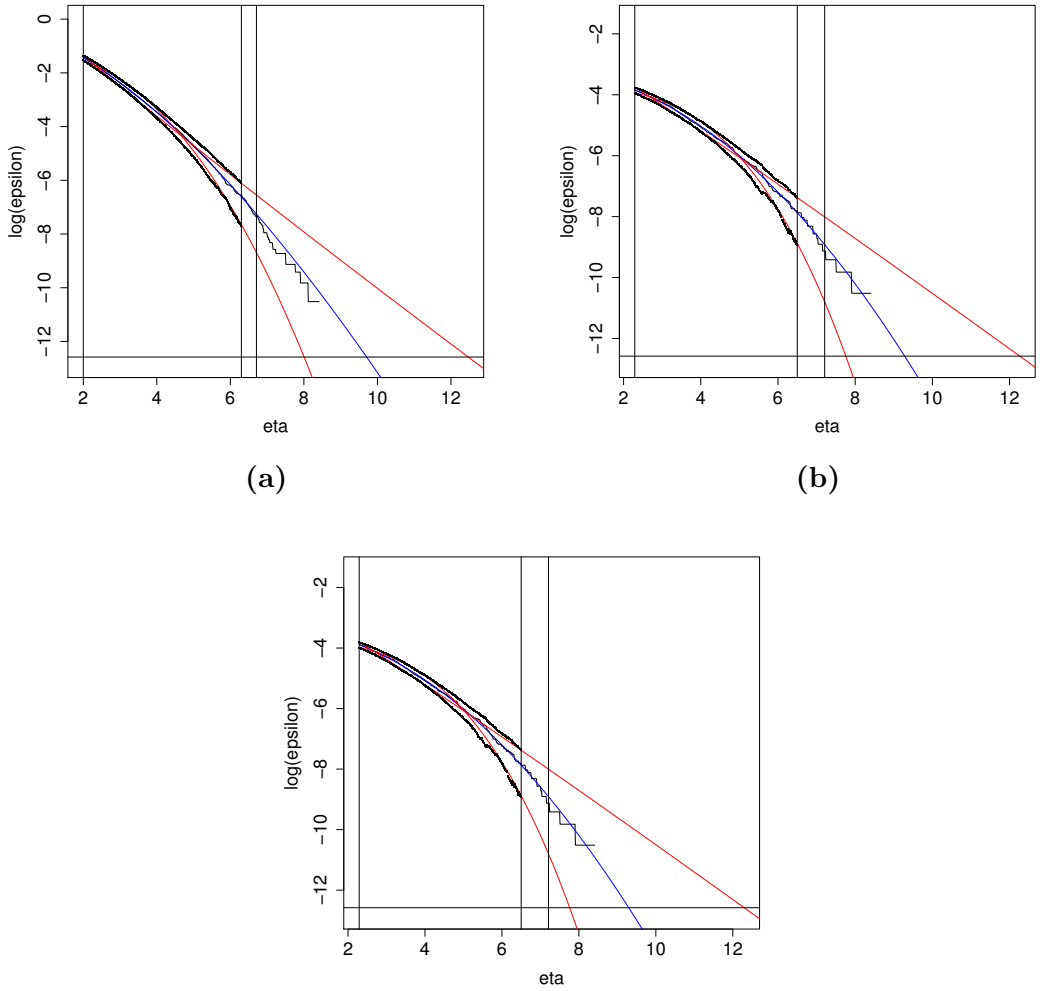


Figure 23: The fitted curve of the Draugen H_S data (in blue) with confidence bands.

during the 12 years and 8 months of the Draugen H_S time series was 14.34. For the H_{SWS} and the H_{SSW} data, however, the curve fit abruptly deviates from the $\log \bar{\epsilon}_k(\eta)$ estimates at $\eta = \eta_2$. Surely, the estimates for $\eta > \eta_2$ are uncertain, but we should not expect those estimates to be very far away from the fitted curve. For the H_{SWS} data, the fitted curve is found to be situated somewhat above the curve of estimates. This seems to indicate that our return level estimates are too high, and might explain the paradox of the H_{SWS} estimates being higher than the H_S estimates. For $k = 2$, the 100 years return level is 16.51, while the highest recorded wave height in the Draugen H_{SWS} time series was only 14.20. For the H_{SSW} data, the opposite seems to be true. Here, the fitted curve seems to give much lower estimates than it should. Actually, for $k = 2$ $\hat{x}_{100} = 9.21$ is only slightly higher than the greatest observation in the Draugen H_{SSW} time series, which was 9.18.

The results for the Ekofisk data are given in table 13 and in figures 27, 28, and 29. As expected, the return levels of the Ekofisk data are lower than the corresponding Draugen levels. However, we still find slightly larger return levels for the H_{SWS} data than for the H_S data, at least in the cases $k = 1$ and $k = 3$.

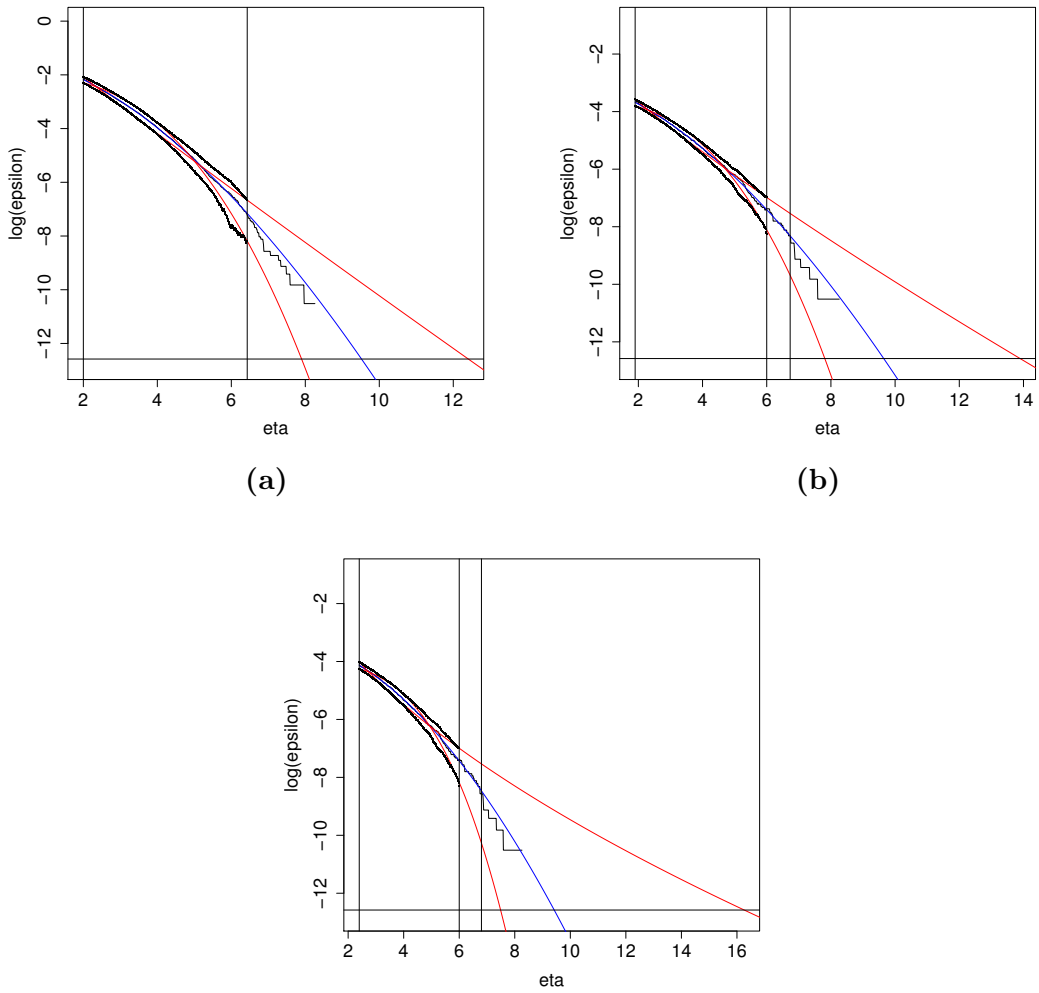
Further, we again find abrupt deviances of the fitted curves from the estimates, but, interestingly, not in the same manner as for the Draugen data. This time, there are large deviations for the H_S data as well as for the H_{SWS} and H_{SSW} data. For the H_S and the H_{SWS} data, we seem to get too low estimates of the return values, while



(a) **(b)**

(c)

Figure 24: Results, Draugen H_S. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.



(a) **(b)**

(c)

Figure 25: Results, Draugen H_{SWS}. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.

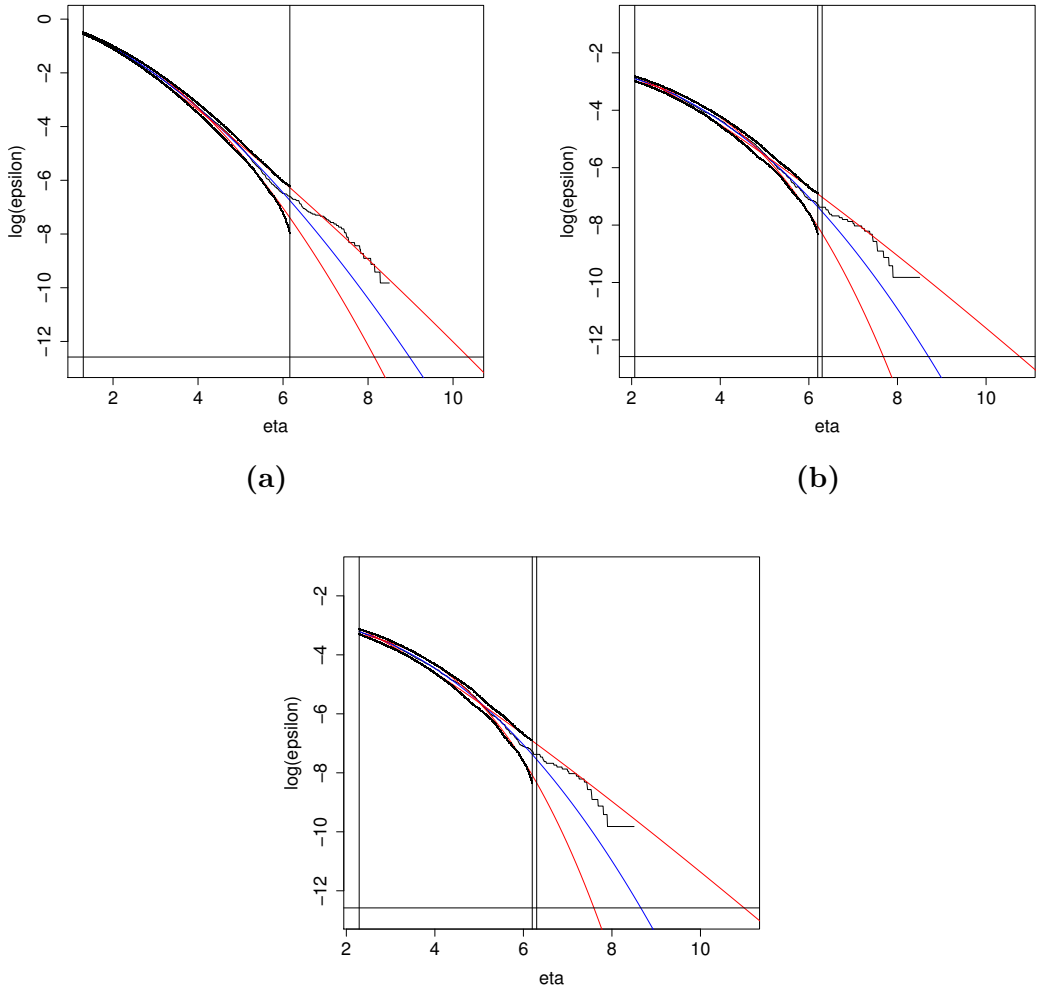


Figure 26: Results, Draugen H_{SSW}. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.

for the H_{SSW} data the estimates seem to be too high. For the H_{SWS} and H_{SSW} data from Draugen, it was the other way around. For $k = 2$, the H_S and H_{SWS} 100 years return level estimates, which are 12.30 and 12.25, respectively, may be compared with the greatest observations from the time series, which are 12.12 and 12.08, respectively. These numbers seem to indicate that the estimated return levels are indeed too low. On the other hand, for $k = 2$, the H_{SSW} 100 years return level estimate $\hat{x}_{100} = 7.33$ is far greater to the greatest among the observations, which is only 5.89.

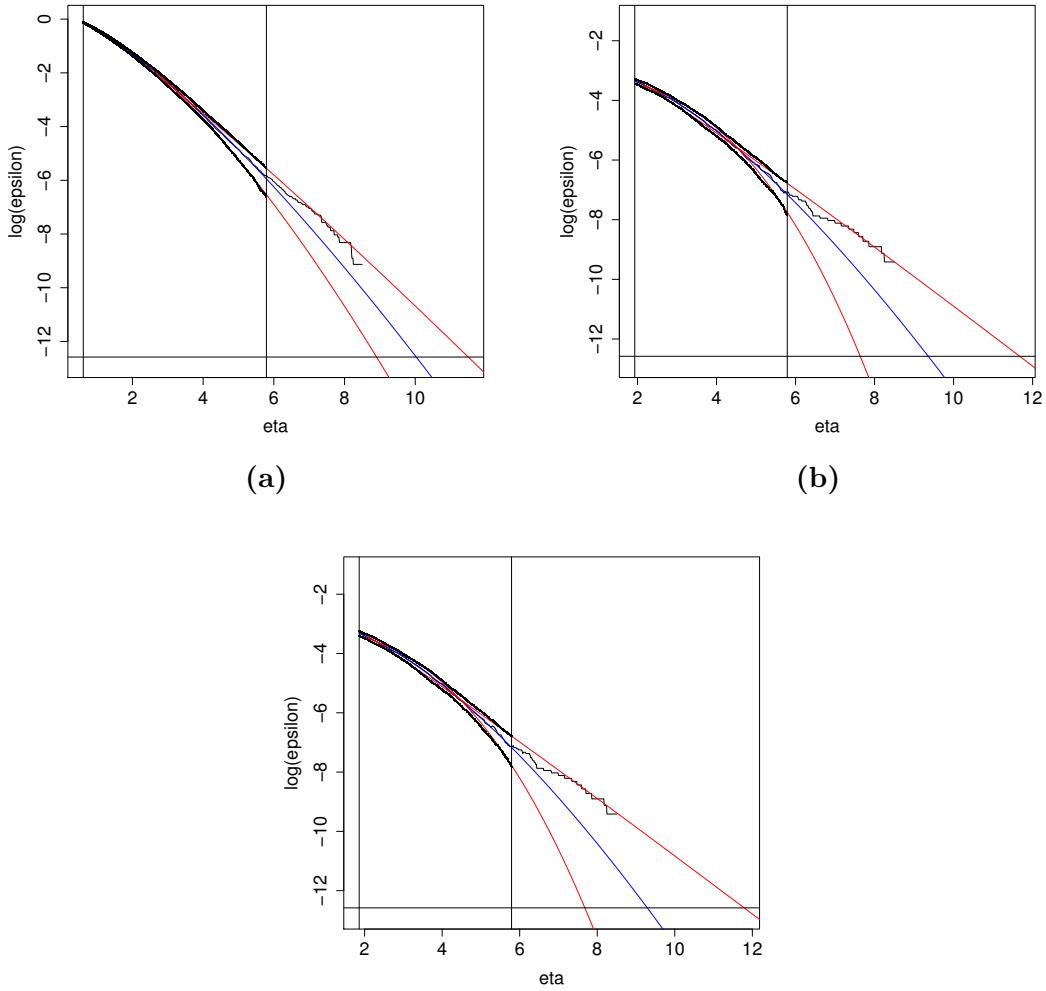
k	η_1	η_2	$\hat{q}$	$\hat{b}$	$\hat{a}$	$\hat{c}$	$\hat{x}_{100}$	$\hat{\text{CI}}(x_{100})$
H_S								
1	0.61	5.79	1.1962	0	0.5866	1.3358	13.1804	(11.7054,15.1149)
2	1.93	5.79	0.0392	1.5575	0.5038	1.4215	12.2999	(10.0372,15.3433)
4	1.86	5.79	0.0470	1.0845	0.3969	1.5097	12.2110	(10.0941,15.4945)
H_{SWS}								
1	0.96	5.37	0.4755	0.6065	0.6941	1.3127	13.1871	(11.0991,15.5892)
2	1.44	5.27	0.0814	0.0800	0.3599	1.5530	12.2530	(10.2069,15.8445)
3	1.44	5.27	0.0725	0.1473	0.3556	1.5597	12.2385	(10.1477,15.8389)
H_{SSW}								
1	1.16	8.10	0.9081	0.9396	0.7438	1.2441	6.4499	(5.8493,6.9785)
2	0.99	5.79	0.3699	0	0.5683	1.2122	7.3256	(6.3618,8.7155)
3	1.41	5.86	0.2514	0.0055	0.4405	1.3157	7.1274	(6.1391,8.6694)

Table 13: Results for the Ekofisk ocean wave data.

Clearly, we should expect the fitted curve to deviate from the estimates for large η, where the uncertainty of the estimates is large. However, such massive deviations as can be found in these plots cannot be explained by uncertainty alone. Our fitted curves do fit the estimates very well for smaller values of η. But the large deviations for $\eta > \eta_2$ seem to indicate that the distributions of the wave heights are inhomogeneous. If that is the case, the return level estimates found by the AER method are invalid, since they are found by extrapolation of the fitted curve $\hat{f}(\eta), \eta_1 \leq \eta \leq \eta_2$ to higher levels of η.

4.4 Discussion

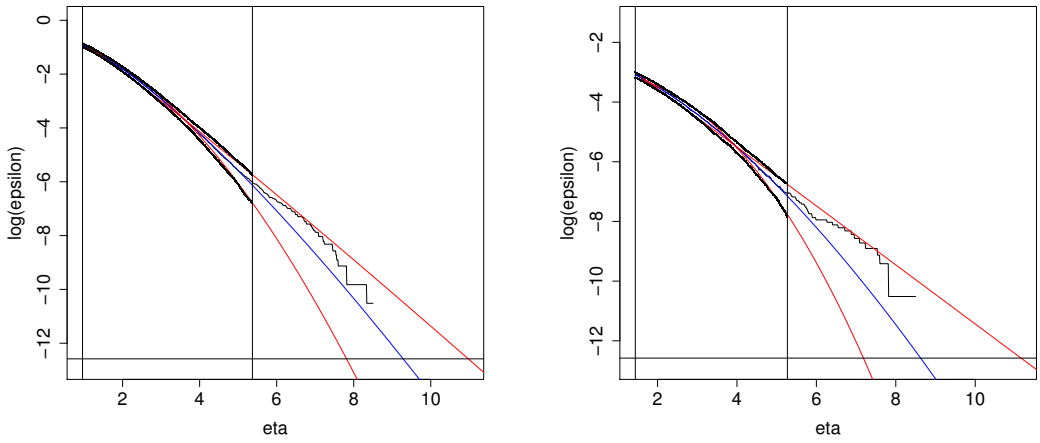
Comparing the different results, we find that the values of the constants vary greatly, even when we might suspect them to be similar. $\hat{b}$ is close to 0 for the Draugen H_S data with $k = 2$, while it is 1.56 for the corresponding Ekofisk data. And the Ekofisk H_S data give almost the same return level estimates for $k = 2$ and the $k = 3$, but in the latter case $\hat{b}$ is 1.08, as opposed to 1.55 in the former case. The fact that we can have such variations in the values of the constants, and still get similar results, seems to indicate that the method is quite robust concerning the estimation of those constants. There may be several combinations of the four constants that yield almost the same return level estimates. The plots in figure 12 also indicates a certain robustness with respect to the choices of η_1 and η_2.



(a) **(b)**

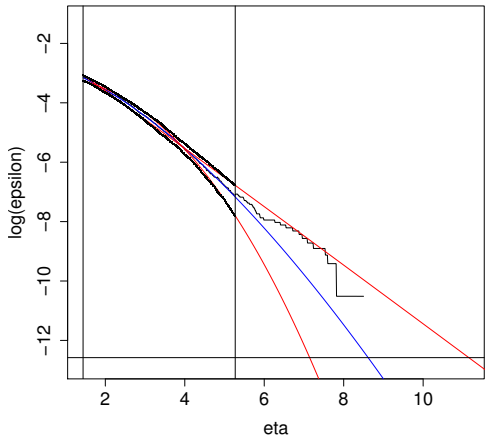
(c)

Figure 27: Results, Ekofisk H_S. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.



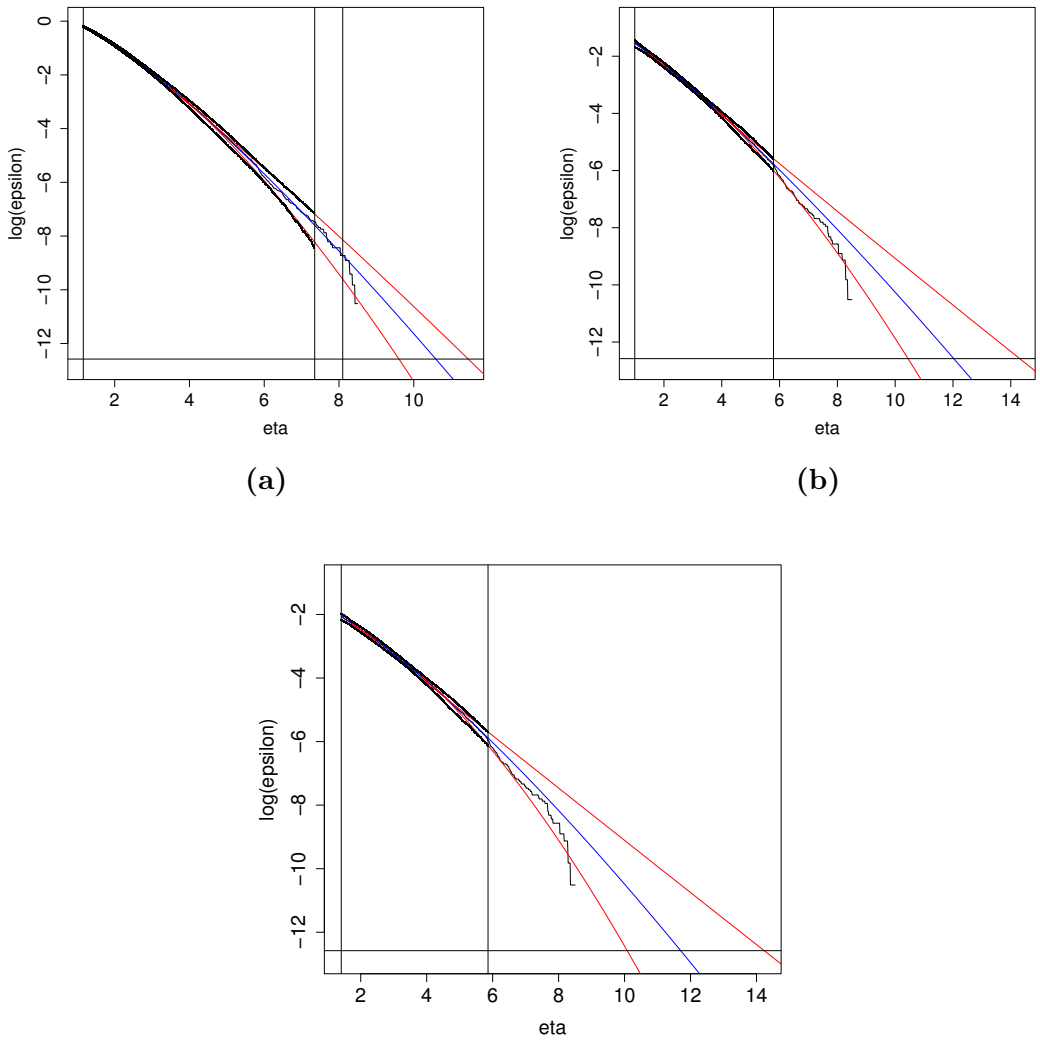
(a)

(b)



(c)

Figure 28: Results, Ekofisk H_{SWS}. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.



(a) **(b)**

(c)

Figure 29: Results, Ekofisk H_{SSW}. **(a)** $k = 1$, **(b)** $k = 2$, **(c)** $k = 3$.

To investigate the robustness with respect to the constants, we try to find out how the error of the log plot fit is dependant upon the choice of constants, particularly the choice of q and b. We do this by varying the values of q and b, while a and c are calculated from the straight line in the log log-log plot. These four constants constitute a fit, the goodness of which is measured by the error given in (42). Again, we use the $w_1(\eta)$ weights formula of (43). As an example we choose the Ekofisk H_{SSW} data with $k = 2$, $\eta_1 = 0.99$, and $\eta_2 = 5.79$, the case that corresponds to the last line but one in table 13.

Now, plotting the error against q and b, we get the plot of figure 30. The smallest error is to be found at the point where $b = 0$ and $q \approx 0.37$. However, the area of the smallest error is seen to be very flat, a fact which suggests that there are many combinations of q and b that will give an error almost as small as the optimal one. Indeed, the numerical method we use to find the optimal value may even have problems finding it. At least is seems advisable to use a small tolerance limit.

This seems to confirm what we have mentioned before, that the choice of the constants is not important *per se*, as long as we get a good fit in the log plot. Indeed, that is all we are interested in. Even if we are not able to find the *optimal* combination of the constants q, b, a, and c, we will have a good estimate of the return level as long as the fitted curve $\hat{f}(\eta)$ follows the estimates $\hat{\epsilon}_k(\eta)$ reasonably well. This is clearly a strength. We may compare this robustness of the AER method against the sensitivity of the POT method concerning its dependency upon the very largest values among the data. Removing only one or two of the greatest data points from the POT analysis may significantly alter the estimates of the return levels, while this will not affect the AER analysis at all, since it leaves the very greatest data out of the analysis.

The robustness of the AER method is fortunate, as the method to estimate the return levels is still, to a certain degree, dependent on human judgement, especially when selecting the best fit after step 1 of the proposed procedure. But given the robustness of the method, we know that although we may not choose the optimal values, our estimates may still be fairly good. As has been pointed out already, the quality of the results can actually be judged by the goodness of the fit. This gives us a criterion for judging the resulting estimates. On the other hand, when using the POT method, it may often be difficult to discern if the fitted extreme value distribution is appropriate or not. In fact, as was explained in section 2.2, it is even impossible to know if the use of the POT method is justifiable at all.

In order to obtain optimal results with the AER method, it was seen to be important to use all available knowledge about the phenomenon under study. Especially, we should try to use our knowledge to find bounds for the constants. For example, our estimates of the Alta wind speed data were much more precise, in the sense that the confidence intervals of the return levels were shorter, when we limited the possible values of the constant c. A lack of such information will make the procedure more difficult to use, as we will have no lower bound on the b values, and must find such a bound from trial and error. Further, the results will be more uncertain, as the confidence intervals will be broader than they need to be. Finally, as was seen from the ocean wave data, it

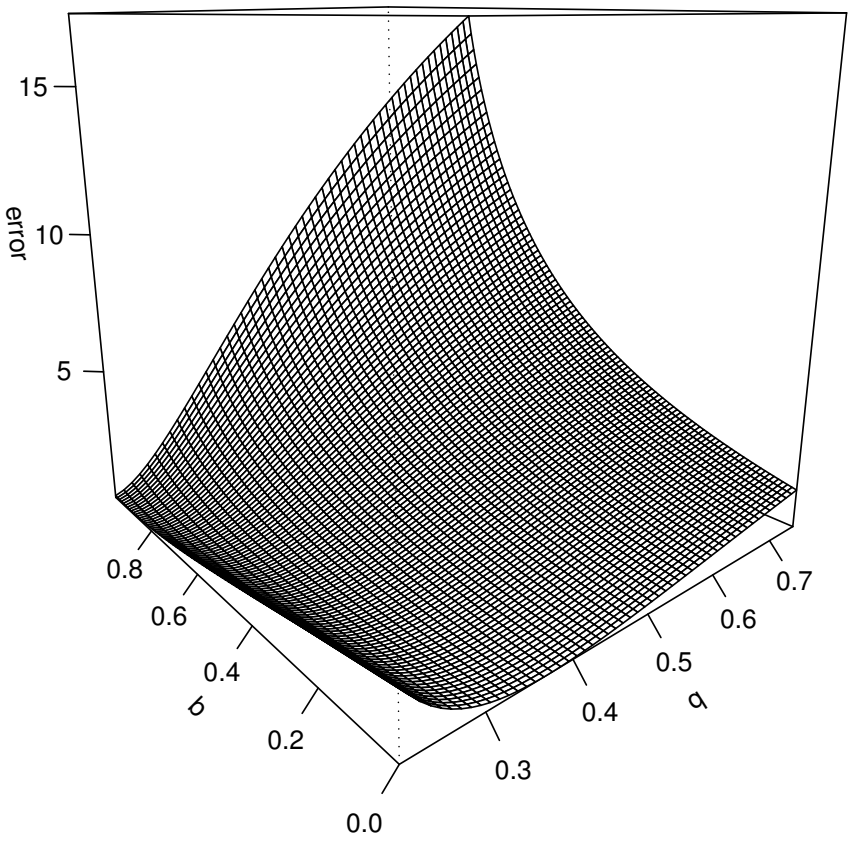


Figure 30: Plot of the error in 42 with different values of q and b, for the Ekofisk H_{SSW} data with $k = 2$.

is very important to ascertain that the AER method can actually be used on the data under study. Especially, we must be sure that the underlying distribution is actually homogenous, so that the extrapolation from known to unknown levels is valid. Drastic deviations of the estimates from the fit for η values larger than η_2 may indicate that this is actually not the case.

In general it can be said that although we have tried to make the proposed procedure as automatic as possible, human judgement, knowledge of the phenomenon under study, and experience in using the method are still crucial for obtaining good results.

5 Conclusion

The AER method seems to be able to give satisfactory results, as long as the data under study belong to a homogeneous distribution. By the proposed procedure, it is possible to find return level estimates with corresponding confidence intervals more or less automatically on a computer. Still, some of the most crucial parts of the analysis are left to human judgement, such as the choice of the η_1 and η_2 values. On the other hand, the method is robust against small errors in the estimated constants, and in the choice of η_1 and η_2 values.

The method seems to have some advantages over the POT method, in that is does not rely on the dubious assumption of asymptotic data. Instead, a wider range of the data can be used in the analysis. Hence, the AER is less dependent upon a few data points in the very tail of the distribution. Evaluation of the results is also made easier, since the quality of the return level estimates is connected to the goodness of a curve fit.

References

[1] Private correspondence with Øyvind Breivik at the Norwegian Meteorological Institute, who provided the ocean wave height data.

[2] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

[3] J. Galambos and N. Macri. Classical Extreme Value Model and Prediction of Extreme Winds. *Journal of Structural Engineering, ASCE*, 125(7):792–794, 1999.

[4] L. H. Holthuijsen. *Waves in Oceanic and Coastal Waters*. Cambridge University Press, Cambridge, 2007.

[5] S. R. Massel. *Ocean Surface Waves: Their Physics and Prediction*. World Scientific, Singapore, 1996.

[6] *Matlab: Optimization Toolbox^{TM} 4. User's Guide*. `http://www.mathworks.com/access/helpdesk/help/pdf_doc/optim/optim_tb.pdf`. Visited June 10, 2008.

[7] Om data og datakvalitet i databasen (in Norwegian). From the web pages of the Norwegian Meteorological Institute. `http://met.no/Meteorologi/A_male_varet/Observasjoner_fra_land/Dette_blir_malt/Kvaliteten_pa_observasjonene/`. Visited June 3, 2008.

[8] eKlima. "External access to climatedata [sic] from Norwegian Meteorological institute". `http://eklima.met.no`. Visited June 3, 2008.

[9] A. Naess. Technical Note: On the long-term statistics of extremes. *Applied Ocean Research*, 6(4):227–228, 1984.

[10] A. Naess. Estimation of Long Return Period Design Values for Wind Speeds. *Journal of Engineering Mechanics, ASCE*, 124(3):252–259, 1998.

[11] A. Naess and O. Gaidai. Estimation of extreme values from sampled time series. *Structural Safety*. To appear.

[12] A. Naess and O. Gaidai. Monte Carlo Methods for Estimating the Extreme Response of Dynamical Systems. *Journal of Engineering Mechanics, ASCE*, 134(8), 2008.

[13] A. Naess, O. Gaidai, and S. Haver. Efficient estimation of extreme response of drag-dominated offshore structures by Monte Carlo simulation. *Ocean Engineering*, 34(16):2188–2197, 2007.

[14] M. Prevosto, H. E. Krogstad, and A. Robin. Probability distributions for maximum wave and crest heights. *Coastal Engineering*, 40(4):329–360, 2000.

A R Code

The following R code is included to illustrate a practical implementation of step 1, as described in section 2.6.

```
linearfunc=function(eta,epsilon,sigma,m,N,dq,db,eta1low,eta1high,eta2low,
    eta2high,neta1,neta2,blow,qhigh)
  {
    ##################################################################

    # INPUT DATA
    # eta - A vector of levels, on a non-dimensional scale.
    # epsilon - A vector of average exceedance rates corresponding to the
        different eta levels.
    # sigma - The standard deviation of the original observations. Will be
        used to transform the return levels back to the original scale.
    # m - The number of periods for which to find the return levels.
    # N - The number of observations per time unit.
    # dq - step length in the discretization of the possible q values.
    # db - step length in the discretization of the possible b values.
    # eta1low - The lower bound of the possible eta_1 values.
    # eta1high - The upper bound of the possible eta_1 values.
    # eta2low - The lower bound of the possible eta_2 values.
    # eta2high - The upper bound of the possible eta_2 values.
    # neta1 - The number of possible eta_1 values to be used.
    # neta2 - The number of possible eta_2 values to be used.
    # blow - The lower bound of the b values.
    # qhigh - The upper bound of the q values.

    # OUTPUT DATA
    # tab - A matrix, containing 9 columns, of which column
    # number (1) contains L values, (2) Delta values, (3) eta1 values,
    # (4) eta2 values, (5) q values, (6) b values, (7) a values,
    # (8) c values, (9) return level values.

    ##################################################################

    # First, we find out which eta values are possible eta1 and eta2
    # values. If the numbers neta1 and neta2 of eta values to be used
    # are smaller than the number of possible values, we pick neta1
    # and neta2 eta values distributed regularly throughout the possible
    # eta_1 and eta_2 values, respectively. Otherwise, we use all the
    # possible values. This procedure results in two new vectors eta1
    # and eta2 containing eta_1 and eta_2 values to be used in the
    # analysis. Vectors epsilon1 and epsilon2 with corresponding epsilon
    # values are also made.

    selection1=which(eta>0 & eta>=min(eta[eta>=eta1low]) & eta<=max(eta[eta
        <=eta1high]))
    selection2=which(eta>=min(eta[eta>=eta2low]) & eta<=max(eta[eta<=
        eta2high]))

    if (length(selection1)>neta1)
```

```
    {
45
      # We select neta1 numbers distributed regularly throughout
47    # (eta1low,eta1high). For each of these numbers we select the eta
      # element which is closest to that number. Hence, we will have
49    # possible eta_1 values distributed more or less regularly
      # throughout (eta1low,eta1high). The vector eta1 will contain
51    # the eta values to be used in the analysis.

53    sequence1=seq(eta1low,eta1high,length=neta1)
      eta.selection1=eta[selection1]
55    select.eta1=matrix(nrow=neta1)

57    for (i in 1:length(sequence1))
        {
59        distance1=abs(eta.selection1-sequence1[i])
          eta.i=union(eta.selection1[distance1==min(distance1)])
61        select.eta1[i]=which(eta==eta.i)
        }
63    eta1=eta[select.eta1]
      epsilon1=epsilon[select.eta1]
65  }
  else
67  {
      eta1=eta[selection1]
69    epsilon1=epsilon[selection1]
      neta1=length(selection1)
71  }

73  if (length(selection2)>neta2)
    {
75    sequence2=seq(eta2low,eta2high,length=neta2)
      eta.selection2=eta[selection2]
77    select.eta2=matrix(nrow=neta2)

79    for (i in 1:length(sequence2))
        {
81        distance2=abs(eta.selection2-sequence2[i])
          eta.i=union(eta.selection2[distance2==min(distance2)])
83        select.eta2[i]=which(eta==eta.i)
        }
85
      eta2=eta[select.eta2]
87    epsilon2=epsilon[select.eta2]
    }
89  else
    {
91    eta2=eta[selection2]
      epsilon2=epsilon[selection2]
93    neta2=length(selection2)
    }
95
  # The possible L values are calculated, and the table tab is
```

```
97    # made, with the rows sorted by the L values. The Delta values are
      # set to infinity to begin with.
99
      diffs=matrix(nrow=length(neta1*neta2))
101   for (i in 1:neta1)
        {
103       for (j in 1:neta2)
            {
105           diffs[(i-1)*neta2+j]=eta2[j]-eta1[i]
            }
107     }

109   sorted=sort(union(diffs,diffs))
      tab=matrix(nrow=length(sorted),ncol=9)
111   tab[,1]=sorted;tab[,2]=Inf

113   # All the combinations of the elements of the  eta1 and eta2 vectors
      # are run through, beginning with the largest eta1 element and the
115   # smallest eta2 element. index1 keeps track of the eta1 elements,
      # while index2 keeps track of the eta2 elements.
117
      index1=neta1
119   while (index1>=1)
        {
121       index2=1
          while (index2<=neta2)
123         {

125           # For each combination, we find the correct L and the correct
              # row rw in the table. Then the possible q and b values are
127           # discretized as vectors Q and B, respectively.

129           differ=eta2[index2]-eta1[index1]
              rw=which(tab[,1]==differ)
131           Q=seq(max(epsilon[eta>=eta1[index1]])+0.000001,max(qhigh,max(
                  epsilon[eta>=eta1[index1]])+0.000001),by=dq)
              B=seq(blow,max(eta1[index1]-0.000001,0),by=db)
133           i=1

135           # We run through all combinations of q and b values.

137           while (i<=length(Q))
                {
139               j=1
                  while (j<=length(B))
141                 {

143                   # For each combination, we investigate beforehand if
                      # the combination can possibly give a smaller Delta
145                   # value for the current L value or not. If not, we skip
                      # the linear fitting.
147
```

```
                    L1=(log(abs(log(epsilon2[index2]/Q[i])))-log(abs(log(
                        epsilon1[index1]/Q[i]))))/(log(eta2[index2]-B[j])-
                        log(eta1[index1]-B[j]))
149                 L2=log(abs(log(epsilon2[index2]/Q[i])))-L1*log(eta2[
                        index2]-B[j])
                    if (max(abs(log(abs(log(epsilon[eta>eta1[index1] & eta<
                        eta2[index2]]/Q[i])))-(L1*log(eta[eta>eta1[index1]
                        & eta<eta2[index2]]-B[j])+L2)))<2*tab[rw,2])
151                   {

153                     # If it is possible to have a smaller Delta value,
                        # we make a linear fit, and calculate the constants
155                     # a (here A) and c (here C) corresponding to the
                        # fit.
157
                        select=which(eta>=eta1[index1] & eta<=eta2[index2])
159                     epsilonsel=epsilon[select]
                        etasel=eta[select]
161                     fit=lm(log(abs(log(epsilonsel/Q[i])))~log(etasel-B[
                            j]))
                        a=exp(as.numeric(fit$coefficients[1]));C=as.numeric
                            (fit$coefficients[2])
163                     line=-log(A)-C*log(etasel-B[j])
                        points=-log(abs(log(epsilonsel/Q[i])))
165                     delta=max(abs(points-line))

167                     # If the new fit actually has a smaller Delta value
                        # than the one in the table, we replace the
169                     # elements in the row with the Delta value, the
                        # eta1 and eta2  elements, the constants and the
171                     # return level of the new fit.

173                     if (delta<tab[rw,2])
                          {
175                         tab[rw,2]=delta;tab[rw,3]=eta1[index1];tab[rw
                                ,4]=eta2[index2]
                            tab[rw,5]=Q[i];tab[rw,6]=B[j];tab[rw,7]=A;tab[
                                rw,8]=C
177                         tab[rw,9]=((-(1/A)*log(-log(1-1/m)/(Q[i]*N)))
                                ^(1/C)+B[j])*sigma
                          }
179                     }
                      j=j+1
181                 }
                  i=i+1
183             }
              index2=index2+1
185         }
          index1=index1-1
187     }
      return(tab)
189 }
```